Large Scale Structure of the Universe:

from simulations to observations

by Santiago Javier Ávila Pérez

PhD Thesis in Theoretical Physics

May 6, 2016

supervised by
Alexander Knebe &
Juan García-Bellido Capdevila

Universidad Autónoma de Madrid

Facultad de Ciencias
Departamento de Física Teórica

&

Instituto de Física Teórica (UAM-CSIC)

A mi familia,
que luchó por verme llegar aquí.

La Estructura a Gran Escala del Universo: Simulando las Observaciones

Authorship

This PhD thesis was authored by Santiago Javier Ávila Pérez.

Chapter 1 is based on the peer-reviewed article [1]. It additionally contains a review on the N-Body simulation field (Section 1.1).

Chapter 2 is based on the peer-reviewed article [2]. It also includes a summary of the results in the peer-reviewed article [3] (Section 2.6).

Chapter 3 is unpublished work done within the Large Scale Structure working group of the Dark Energy Survey collaboration.

[1] *SUSSING MERGER TREES: the influence of the halo finder*
Avila S.; Knebe A.; Pearce F.R.; Schneider A.; Srisawat C; Thomas P.A.; Behroozi P.; Elahi P.J.; Han J.; Mao Y.; Onions J.; Rodriguez-Gomez V. and Tweed D.
2014 MNRAS 441 p. 3488-3501

[2] *HALOGEN: a tool for fast generation of mock halo catalogues*
Avila S., Murray S.G., Knebe A., Power C., Robotham A. and Garca-Bellido J.
2015 MNRAS 450 p. 1856-1867

[3] *nIFTy Cosmology: galaxy/halo mock catalogue comparison project on clustering statistics*
Chuang C.-H., Zhao C., Prada F., Munari E., Avila S., Itzard, A., et al. Kitaura F.S., Monaco, P., Murray S., Knebe A., Scoccola C.G., Yepes G., Garcia-Bellido J., Marin F., Muller V., et al.
2015 MNRAS 452 p. 686-700

Contents

Prólogo

El Modelo Cosmológico Estándar: ΛCDM

El campo de la Cosmología ha llegado a un modelo de concordancia capaz de conciliar todos los experimentos cosmológicos: ΛCDM. Este modelo se basa en los principios de la teoría del Big Bang, que explica la expansión del Universo a través de la ecuación de Friedman. Ésta puede ser expresada como:

$$H(a) = \frac{\dot{a}}{a} = H_0 \sqrt{(\Omega_{\rm c} + \Omega_{\rm b})\, a^{-3} + \Omega_{\rm rad} a^{-4} + \Omega_{\rm k} a^{-2} + \Omega_{\rm DE} a^{-3(1+w)}} \qquad (1)$$

donde Ω_i representa el parámetro de densidad de la especie i. Es decir, el cociente entre la densidad ρ_i y la densidad crítica $\rho_{crit} = \frac{3H_0^2}{8\pi G}$ en la época actual.

Los experimentos encuentran una componente de materia ordinaria (o bariónica) de $\Omega_{\rm b} = 0.049$, y una pequeña componente de radiación (fotones y neutrinos) de $\Omega_{\rm rad} = 9 \cdot 10^{-5}$. La novedad con respecto a la *antigua* teoría del Big Bang es el efecto dominante de dos nuevas especies: la Materia Oscura Fría ($\Omega_{\rm c} = 0.266$) y la Energía Oscura ($\Omega_{\rm DE} = 0.685$). En el Modelo Estándar ΛCDM, la componente de curvatura es insignificante ($\Omega_{\rm k} = 0$), y la ecuación de estado de la Energia Oscura es $w = -1$, que corresponde a una Constante Cosmológica Λ. La velocidad de expansión actual del Universo es $H_0 = 67.3(km/s)/Mpc$. Todos los valores citados proceden de [4].

La Materia Oscura Fría (CDM, en adelante todas las siglas tienen su origen en la nomenclatura en lengua inglesa) es indistinguible de la materia ordinaria en su comportamiento gravitatorio y, por tanto, la forma en que afecta a la expansión del Universo (Ecuación 1). Sin embargo, es necesaria una componente de materia acol-

isional y no bariónica para explicar las estructuras que encontramos en el Universo. La presencia de CDM se ha determinado a través de mediciones de las curvas de velocidad de las galaxias, la velocidad de dispersión de los cúmulos de galaxias, el efecto de lentes gravitacionales y el fondo de radiación cósmico (CMB). No obstante, aún no se ha producido una detección directa que pueda esclarecer su naturaleza [5].

La Energía Oscura es una fuerza de repulsión introducida en la ecuación de Friedman para explicar la aceleración del Universo en épocas recientes. La naturaleza de esta componente permanece desconocida, aunque el modelo más simple (la Constante Cosmológica) asume que es equivalente a una energía de vacío. Sin embargo, las diferentes teorías cuánticas de campos predicen una energía de vacío entre 60 y 120 órdenes de magnitud superior al valor observado. Por tanto, una extensión de la teoría parece necesaria, y entre las propuestas encontramos teorías modificadas de la gravedad [6].

Desde un punto de vista fenomenológico, la Energía Oscura se puede parametrizar con la ecuación de estado w (el cociente entre la presión y la densidad), necesitando $w < -1/3$ para una expansión acelerada. También podemos parametrizarla con una ecuación de estado que depende del tiempo:

$$w = w_0 + (1 - a)w_a \tag{2}$$

Un tercer ingrediente en este paradigma es la Inflación: una época de expansión acelerada del Universo justo después del Big Bang. Esta teoría es capaz de explicar la planitud ($\Omega_{\rm k} = 0$) del Universo, su homogeneidad a grandes escalas y el origen de las estructuras. Más específicamente, predice un espectro de potencias de las fluctuaciones primordiales cercano a la invariancia de escalas $\mathcal{P}_S(k) \propto k^{n_s - 1}$ [7].

Cabe comentar que ciertas publicaciones o ciertos autores aseguran que se han encontrado evidencias observacionales en contra del Modelo Estándar ΛCDM: la abundancia de cúmulos de características extremas [8, 9], los problemas encontrados en las estructuras a pequeñas escalas (e.g. la escasez de galaxias satélites [10]) o las anomalías encontradas en las grandes escalas del CMB [11]. No obstante, ninguna de esas pruebas son concluyentes y el debate sigue abierto. En el caso del problema de pequeñas escalas, se puede argüir que nuestro conocimiento del efecto de la ma-

teria bariónica a esas escalas es limitado y puede estar detrás de la causa (e.g. [12]). De hecho, con nuevas observaciones se están encontrando soluciones a algunos de los problemas que llevaban tiempo esperando respuesta (e.g. el descubrimiento de nuevas galaxias satélites [13]). Respecto a los otros dos casos, estimar la probabilidad de eventos extraños (la existencia de cúmulos extremos y las peculiaridades del CMB), una vez que sabemos que ocurren, puede ser una tarea complicada y subjetiva. Nuevas estimaciones de otros grupos encuentran que las citadas anomalías son compatibles con ΛCDM [14–16].

La Revolución Cosmológica

Vivimos actualmente una Revolución Cosmológica. En los últimos 20 años la Cosmología Observacional ha evolucionado desde sólo poder estimar el orden de magnitud de la cantidad de materia en el Universo $\Omega_m = \Omega_b + \Omega_c$ o el ritmo de expansión H_0, a medirlos con una precisión por debajo del 2%, a descubrir la existencia de la Energía Oscura, determinar la planitud del Universo, poner cotas a la masa de los neutrinos, encontrar desviaciones de la invariancia de escala perfecta, medir factores de crecimiento, las distorsiones en el espacio de redshift, las oscilaciones acústicas de bariones (BAO), etc.

En la Figura 1 presento una visión más personal de esta Revolución Cosmológica. Los dos paneles superiores, obtenidos de la recopilacilación Union2.1 [1] [17], representaban el conocimiento que la comunidad tenía de la Cosmología cuando empecé el doctorado en 2012. Pero durante los últimos 4 años han quedado obsoletos, dado que los resultados del experimento Planck[2] [4, 18] han puesto los límites más restrictivos a los parámetros cosmológicos, representados en los paneles centrales. Mientras que los paneles superiores solían mostrarse en cualquier presentación ligeramente relacionada con cosmología, ese rol lo han tomado ahora los datos de Planck. Pero todavía hay mucho trabajo por delante, ya que seguimos avanzando hacia la Cosmología de Precisión. Los dos paneles inferiores muestran predicciones de los límities que se impondrán en la próxima década: en la izquierda con los datos finales del Dark

[1] http://supernova.lbl.gov/union/
[2] http://www.cosmos.esa.int/web/planck

Energy Survey (DES)[3] [19], y en la derecha con los datos del futuro cartografiado Euclid[4] [20]. El principal objetivo de estos dos experimentos es medir con una precisión sin precedentes la ecuación de estado de la Energía Oscura y su variación temporal (Ecuación 2). Esto nos ayudará a comprender la naturaleza de esa misteriosa fuerza que domina la densidad de energía del Universo.

Como hemos visto en la parte superior de la Figura 1, tres tipos principales de experimentos han contribuido en las primeras etapas de la Cosmología de Precisión. A continuación, los repasaré brevemente.

Las Supernovas de tipo Ia (SNIa) son explosiones violentas que pueden ser usadas como *candelas estándares* (objetos cuya luminosidad es conocida) y observadas a distancias cosmológicas. Midiendo su desplazamiento al rojo o redshift z, podemos estudiar la relación entre éste y la distancia de luminosidad:

$$d_L(z) = (1+z)\int_0^z \frac{c\, dz'}{H(z')} \tag{3}$$

que depende fuertemente de los parámteros cosmológicos que determinan la evolución del Universo ([21, 22]).

Hacia el final del pasado siglo (1998-1999), los equipos de High-Z Supernova Search Team[5] y Supernova Cosmology Project[6] midieron $d_L(z)$, encontrando que –al contrario de lo que se esperaba– la expansión del Universo se estaba acelerando [23, 24]. Esto suposo la primera de una serie de evidencias acerca de la existecia de la Energía Oscura.

Oscilaciones Acústicas de Bariones (BAO). Al principio, el Universo primigéneo consistía en un plasma a altas temperaturas donde la materia bariónica estaba ionizada e interactuaba muy fuertemente con los fotones. Las perturbaciones iniciales, que son sembradas por la Inflación, se propagan por el plasma a través de

[3] http://www.darkenergysurvey.org/es
[4] http://www.euclid-ec.org/
[5] https://www.cfa.harvard.edu/supernova/public.html
[6] http://www-supernova.lbl.gov/

ondas de sonido causadas por los gradientes de presión. En un momento dado, el Universo se vuelve neutro (*recombination*), y poco después se produce el desacoplo (*decoupling*): las materia bariónica deja de interactuar con los fotones. En ese instante, $a_{\rm dec}$, las oscilaciones se *congelan* y dejan impresa la escala del horizonte del sonido $\chi_{\rm BAO}$ en la distribución de materia [25, 26]:

$$\chi_{\rm BAO} = \frac{c}{\sqrt{3}} \int_0^{a_{dec}} \frac{da}{a^2 H(a) \sqrt{1 + 3\Omega_b/(4\Omega_\gamma)}} \tag{4}$$

Esta escala se puede observar a diferentes tiempos cósmicos en la distribución de galaxias (u otro identificador de la materia) como una protuberancia en la función de correlación a grandes escalas. Finalmente, podemos usar esta escala como una *regla estándar* (un objecto cuyo tamaño es conocido) para medir la relación entre la distancia angular y el redshift:

$$d_A(z) = \frac{1}{1+z} \int_0^z \frac{c\, dz'}{H(z')} \tag{5}$$

que se relaciona con la Ecuación 3 a través de $d_M(z) = d_A(z) \cdot (1+z) = d_L(z)/(1+z)$. En la Figura 2 se muestra conjuntamente la señal del BAO y de las SNIa.

A pesar de que entender profundamente la física altamente no líneal asociada a la formación de estructuras y su relación con los observables no es tarea fácil, la física que determina $\chi_{\rm BAO}$ se basa en princios mucho más sencillos. De este modo las señales del BAO medidas en la Estructura a Gran Escala (LSS) se convirtieron pronto en una fuente muy provechosa para determinar los parámetros cosmológicos. El BAO fue detectado por primera vez en la distribución de galaxias por las colaboraciones 2dFGRS[7] [27] y SDSS[8] [28]. Más adelante otras medidas más precisas fueron establecidas por 6dFGS[9] [29], WiggleZ[10] [30] y BOSS[11] [31]. Además de estas medidas, BOSS midió el BAO a través de los llamados bosques de Lyman-α: reconstruyendo la distribución tridimensional de la posición de las balsas de hidrógeno

[7] http://www.2dfgrs.net/
[8] http://www.sdss.org/
[9] http://www.6dfgs.net/
[10] http://wigglez.swin.edu.au/site/
[11] http://Cosmology.lbl.gov/BOSS/

neutro intergaláctico que dejan líneas de absorción en los espectros de cuásares lejanos [32, 33].

Fondo Cósmico de Radiación (CMB). Tras el desacoplo de fotones y bariones, la luz puede viajar libre, dejando un baño de fotones en el Universo que vemos a día de hoy a la temperatura de $T_{\rm CMB} = 2.73K$: el CMB. Ésta era una de las predicciones clave de la teoría del Big Bang, que fue confirmada experimentalmente tras el descubrimiento de Penzias y Wilson [34].

Esta temperatura es isotrópica en una fracción $\sim 1/10^5$, pero existen pequeñas fluctuaciones originadas durante la Inflación, que más tarde se propagan en el plasma de luz y materia, como se explicó anteriormente. Después del desacoplo, los rayos de luz son ligeramente desviados por la Estructura a Gran Escala (CMB lensing [35]) y la energía de los fotones se ve modificada por el colapso de las estructuras (efecto ISW, [36]). A día de hoy, podemos medir esas fluctuaciones, que contienen información acerca de esas tres épocas. El efecto más visible es, precisamente, el pico del BAO. Sin embargo, en el resto de esta tesis el BAO se referirá a la señal que se queda grabada en la distribución de materia a menos que se indique lo contrario.

Entre los experimentos del CMB cabe destacar los tres satélites: COBE[12] [37], que detectó por primera vez las anisotropías del CMB; WMAP[13] [38], que midió los tres primeros picos del espectro de potencias, y Planck [11], que completó las medidas del espectro hasta escalas de $\sim 0.07°$.

Cosmología de Precisión con la Estructura a Gran Escala

En la actualidad, la mayor fuente de información para limitar el rango de los parámetros cosmológicos procede del CMB. Esto es en parte debido a que la física del CMB se puede entender y modelizar fácilemtne con teoría líneal de perturbaciones. Por ello, este campo se desarrolló muy rápido y fue pionero en la Cosmología de Precisión. Sin embargo, la extracción de inforamación cosmológica del CMB ha alcanzado ya su máximo y se centra ahora en estudios de orden superior, que son

[12] http://lambda.gsfc.nasa.gov/product/cobe/
[13] http://map.gsfc.nasa.gov/

más sutiles (polarización, distorsión espectral, etc.)

En el otro extremo, cada vez entendemos mejor el campo de la Estructura a Gran Escala (LSS). Esto se debe a un mejor control de los errores sistemáticos e instrumentos más precisos y especializados, pero también a un mayor entendimiento de la astrofísica de los observables y un mayor conocimiento de la cosmología subyacente.

Además, mientras el CMB sólo nos aporta un mapa de dos dimensiones con información proviniente principalmente de la física previa al desacoplo y durante éste; LSS explora el espacio tridimensional a tiempos más recientes, cuando la influencia de la Energía Oscura empeza a ser importante. Se puede explorar LSS con diferentes identificadores de la materia: galaxias, cuásares, cúmulos de galaxias, bosques Ly-α, gas HI, etc.

Anteriormente, vimos el potencial de los cartografiados de galaxias para medir el BAO. Pero con un mejor entendimiento de LSS y los datos adecuados, los cartografiados de galaxias pueden poner nuevas cotas sin precedentes en los parámetros cosmológicos. Cuando se trata de determinar la naturaleza de la Energía Oscura y de distinguir entre la Constante Cosmológica y un modelo de gravedad modificada, toda información es útil, ya que los modelos alternativos de gravedad pueden mostrar su signo diferenciador en muchos tipos de medidas diferentes.

Entre las metas de los cartografiados venideros se encuentra medir el espectro de potencias completo para limitar el rango de n_s, hacer tomografía 3D del Universo con *weak lensing*, medir la función de correlación a 3 puntos para encontrar no-gausianidades, determinar los factores de crecimiento a diferentes redshifts, estimar la función de masas de los halos (especialmente el final de la distribución), etc. [19, 20]

Para lograr Cosmología de Precisión con observaciones de la Estructura a Gran Escala, necesitamos un modelo con el que contrastar los datos. Aunque existen modelos teóricos de la formación de estructura [39], las predicciones son limitadas, y tenemos que recurrir a simulaciones de N-cuerpos (N-Body), y una serie de herramientas asociadas (véase la Sección 1.1 y las referencias allí citadas). En la Figura 3, vemos una representación de las galaxias observadas en un cartografiado (en azul) y su equivalente a través de simulaciones (en rojo). Vemos muchas similitudes en la dis-

tribución de galaxias, que forman estructuras muy complejas y altamente no lineales: supercúmulos, filamentos, muros y vacíos.

Las herramientas que utilizamos en las simulaciones deben de ser validadas para verificar si son adecuadas para la Cosmología de Precisión. En esas líneas, en el Capítulo 1, evaluamos una serie de *Halo Finders* y *Merger Tree builders*. Los *Halo Finders* son herramientas que se usan para identificar objetos densos o ligados gravitacionalmente –halos– en la distribución de materia oscura de las simulaciones (y que sirven para alojar galaxias). Los *Merger Tree builders* son otro tipo de herramientas, diseñadas para reconstruir la historia de los halos en las simulaciones. Partiendo de la misma simulación de materia oscura, aplicamos diferentes combinaciones de las citadas técnicas para analizar los *Merger Trees* (un esquema con toda la historia de un halo) resultantes. Estudiaremos la edad de los halos, la probabilidad de fusión entre halos y la evolución de su masa, según los resultados extraídos con cada una de las diferentes técnicas.

Además, las medidas de LSS necesitan barras de error y matrices de covariancia que cuantifiquen los errores sistemáticos, la varianza cósmica y su combinación. Para poder estimarlas, no es suficiente con tener una simulación, sino que necesitamos cientos o miles de ellas. Las simulaciones N-Body, son muy costosas en términos de recursos informáticos. Por ello, generar tantas, con los volúmenes requeridos por los cartografiados actuales, y con sufciente resolución de masa, está fuera del alcance de cualquier grupo de investigación, incluso con los supercomputadores más punteros. Por ejemplo, una de las simulaciones que tomaremos como referencia, MICE, fue llevada a cabo en el supercomputador MareNostrum[14] (véase la Sección 2.1 y las referencias allí citadas). Por tanto, necesitamos una nueva generación de herramientas de simulación, capaz de generar catálogos simulados de manera apróximada. En el Capítulo 2, presento HALOGEN, un método diseñado para generar catálogos de halos que presenten una función de correlación a 2 puntos correcta a grandes escalas. Con este método el número de horas de CPU se reduce en un factor $\sim 10^{3-5}$ y la memoria en un factor $\sim 10^{1-2}$ con respecto a una simulación N-Body (Tabla 2.5). Además veremos otros métodos aproximados para la generación de catálogos simulados y una comparación de todos ellos.

[14] https://www.bsc.es/

Para finalizar, en el Capítulo 3, presento la aplicación de HALOGEN en el contexto del análisis de los datos del Dark Energy Survey. En primer lugar, al método le añadimos tres nuevas implementaciones de caracter observacional: la construcción de un cono de luz, la simulación de un redshift fotométrico, y la implementación de un método de ocupación de halos con galaxias, (HOD, véase referencias en la Sección 3.2.3) que es ajustado para reproducir la correlación de las galaxias observadas. Después, generamos una remesa de catálogos de galaxias y demostramos cómo pueden ser (y están siendo) utilizados para: comprender mejor la modelización y la física de las obserservaciones que simulamos; optimizar la metodología del análisis y, por último, calcular barras de error y matrices de covariancia para el análisis de la Estructura a Gran Escala con los datos de DES.

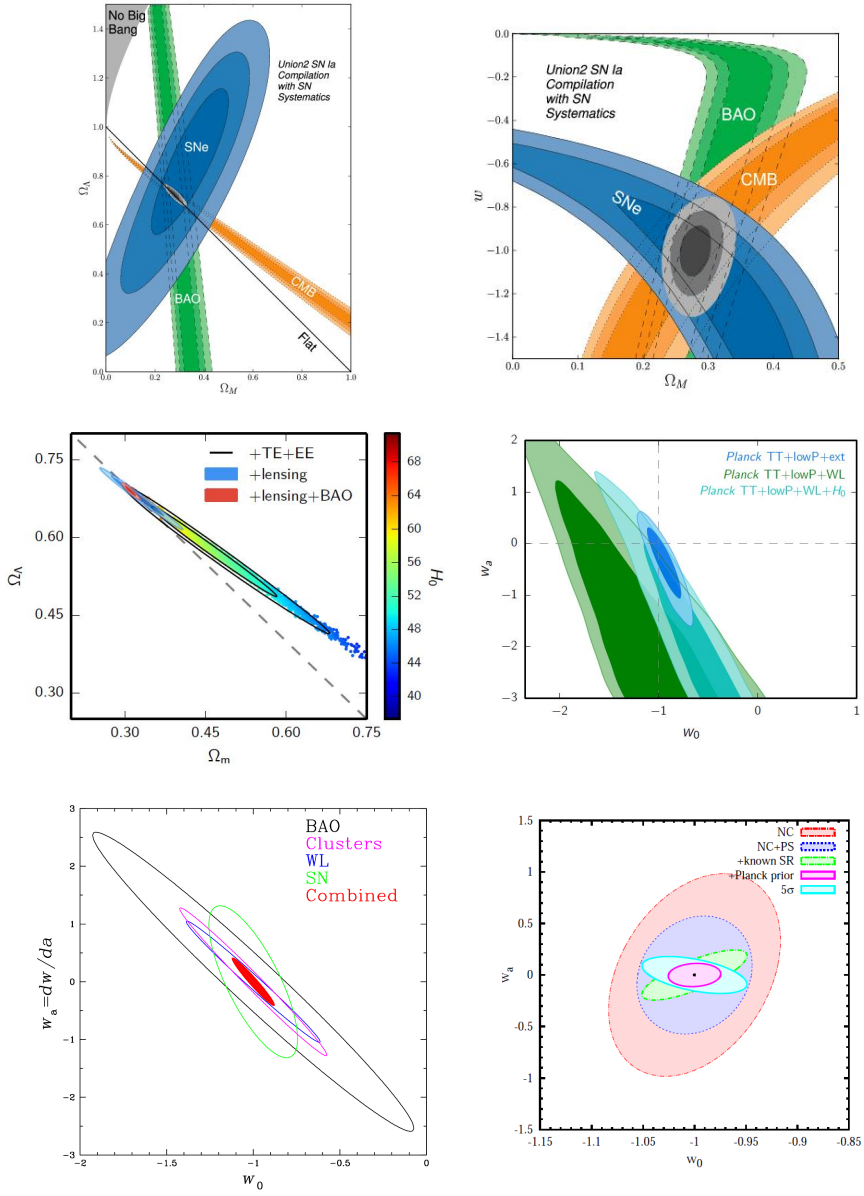

Figure 1: Cosmological Constraints. **Top:** Constraints from Union2 [17] (2010) combining data from SNIa, CMB and BAO. 1, 2 & 3 σ confidence level contours of the density parameter of matter and Dark Energy assuming $w = -1$ (Left), and constraints of the density of matter and the equation of state of Dark Energy assuming a flat Cosmology (Right). **Middle**: Constraints from Planck 2015 release [4]. Contours of the 1 & 2 σ region of the density parameters (Left) and on the time-dependent equation of state of Dark Energy (Right), parametrised by Equation 2. Note, that constraints become strong when combined with other probes (red on the left, blue on the right). **Bottom**: Forecast for DES [40] (Left) and Euclid [41] (Right), 1-σ constraints on the $\{w_0, w_a\}$ plane.

Preface

The Standard Cosmological Model: ΛCDM

A concordance model has been reached in the field of Cosmology, able to reconcile all the cosmological experiments: ΛCDM. This model is based on the principles of the Hot Big Bang theory, that explains the expansion of the Universe through the Friedman equation, which can be expressed as

$$H(a) = \frac{\dot{a}}{a} = H_0 \sqrt{(\Omega_{\rm c} + \Omega_{\rm b})\, a^{-3} + \Omega_{\rm rad} a^{-4} + \Omega_{\rm k} a^{-2} + \Omega_{\rm DE} a^{-3(1+w)}} \qquad (1)$$

with Ω_i representing the density parameter of the species i. This is, the ratio of the density ρ_i to the critical density $\rho_{crit} = \frac{3H_0^2}{8\pi G}$ at the current epoch.

Experiments find a component of ordinary (baryonic) matter of $\Omega_{\rm b} = 0.049$ and a small component of radiation (photons and neutrinos) of $\Omega_{\rm rad} = 9 \cdot 10^{-5}$. The novelty with respect to the *old* Big Bang theory is the dominant effect of two new species: the Cold Dark Matter ($\Omega_{\rm c} = 0.266$) and the Dark Energy ($\Omega_{\rm DE} = 0.685$). In the standard ΛCDM model, the curvature is negligible ($\Omega_{\rm k} = 0$) and the equation of state of Dark Energy is $w = -1$, corresponding to a Cosmological Constant Λ. The measured current expansion rate is $H_0 = 67.3(km/s)/Mpc$ (all quoted values from [4]).

The Cold Dark Matter is indistinguishable from ordinary matter in its gravitational behaviour and, hence, in the way it affects the expansion of the Universe (Equation 1). However, we need a non-baryonic collisionless component to explain the formation of the structures that we find in the Universe. Its presence has been

determined by measurements of rotational curves of galaxies, velocity dispersion of galaxies in galaxy clusters, gravitational lensing and Cosmic Microwave Background (CMB). But we are still missing a direct detection of Dark Matter that can shed light upon its nature [5].

The Dark Energy is a repulsive force introduced in the Friedman equation to account for the acceleration of the Universe at late times. The nature of this component remains unknown, although the simplest model (the Cosmological Constant) assumes it is equivalent to a vacuum energy. However, predictions on the vacuum energy from Quantum Field Theories disagree in $\sim 60 - 120$ orders of magnitude with observations. Hence, extensions of the theory appear necessary amongst which we find modified theories of gravity [6].

From a phenomenological approach, Dark Energy is parametrised by its equation of state w (the pressure to density ratio), needing $w < -1/3$ for an accelerated expansion. It can further be parametrised as a time-dependent equation of state as

$$w = w_0 + (1 - a)w_a \tag{2}$$

A third ingredient in this paradigm is Inflation: an epoch of accelerated expansion just after the Big Bang. It explains the flatness ($\Omega_{\rm k} = 0$) of the Universe, its homogeneity at large scales and the origin of structures. Particularly, inflation predicts a nearly scale invariant power spectrum of primordial fluctuations $\mathcal{P}_S(k) \propto k^{n_s - 1}$ [7].

As a final comment, some experimental evidences against ΛCDM have been claimed: the abundance of extreme clusters [8, 9], problems with structure formation at small scales (e.g. the missing satellite problem [10]) or the large scale anomalies of the CMB [11]. However, none of those evidences were conclusive and the debate remains open. The small scale problems can be argued away by highlighting the little knowledge we have about the effect of baryon physics on those scales (e.g. [12]). In fact, new observations are solving some of the long-lasting challenges (with many more satellites found [13]). As for the other two cases, determining the probability of anomalous events (large scale features in the CMB or abundance of extreme clusters) happening once we know they do occur is not trivial. Estimations from other groups find the anomalies compatible with ΛCDM [14–16].

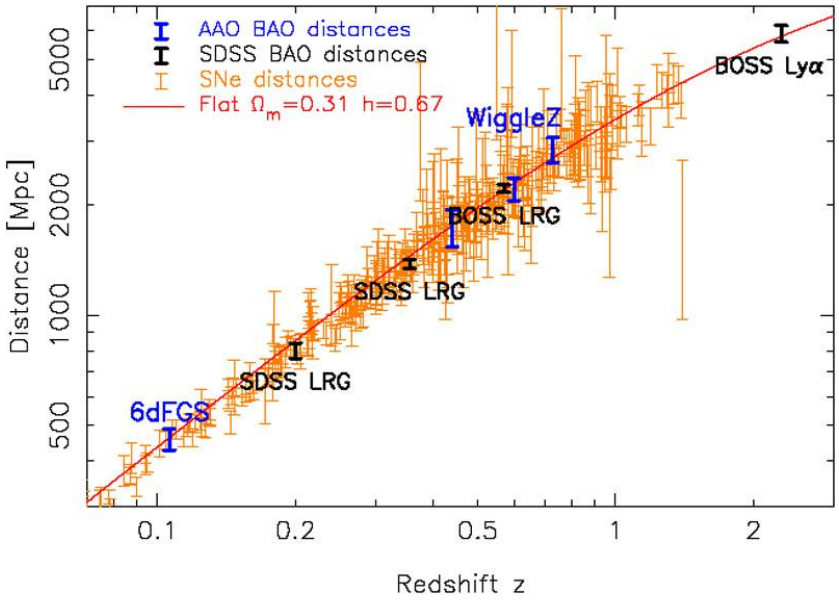

Figure 2: Distance-redshift relation. Compilation of SNIa from [42] (2008) and BAO measurements from [29–31, 33]. Figure from C. Blake in [43].

Cosmological Revolution

We are currently living in a Cosmological Revolution. In the last 20 years Observational Cosmology has evolved from struggling to estimate the amount of matter in the Universe $\Omega_m = \Omega_b + \Omega_c$ or the rate of expansion H_0, to measuring them with a $< 2\%$ accuracy, measuring the flatness of the Universe, discovering the existence of Dark Energy, setting constraints on neutrino masses, finding deviations from perfect scale independence in the primordial power spectrum, measuring growth factors, redshift space distortions, Baryonic Acoustic Oscillations (BAO), etc.

In Figure 1 I present a more personal perspective of this Cosmological Revolution. The two top panels from the Union2.1 compilation[15] [17] represented the knowledge the scientific community had about Cosmology when I started my PhD in 2012. But during the last four years they became outdated, since the results from the Planck experiment[16] [4, 18] set the most stringent constraints of cosmological parameters,

[15] http://supernova.lbl.gov/union/
[16] http://www.cosmos.esa.int/web/planck

represented in the middle panels. The top panes used to appear in any presentation slightly related to Cosmology, nowadays, Planck data have taken over that role. But there is still much work to be done in the future, as we keep advancing towards Precision Cosmology. The two bottom panels show constraints forecast for the next decade: on the left for the completed Dark Energy Survey[17] [19] and on the right for the future survey Euclid[18] [20]. The main target of both of these experiments is to set unprecedented constraints on the time-dependent equation of state of Dark Energy (Equation 2). This will help us understanding the nature of this mysterious force that dominates the energy density of the Universe.

As seen at the top of Figure 1, three main type of experiments contributed to the early stages of Precision Cosmology. I will briefly review them below.

Type Ia Supernova (SNIa) are violent explosions that can be used as standard candles and detected at cosmological distances. Measuring their redshift z, we can study the luminosity distance-redshift relation

$$d_L(z) = (1+z)\int_0^z \frac{c\ dz'}{H(z')} \tag{3}$$

highly dependent on the cosmological parameters that determine the evolution of the late Universe ([21, 22]).

At the end of the last century (1998-1999) the High-Z Supernova Search Team[19] and Supernova Cosmology Project[20] measured $d_L(z)$ determining that instead of decelerating –as it was expected– the expansion of the universe was accelerating [23, 24]. This was the first evidence for Dark Energy.

Baryonic Acoustic Oscillations (BAO). At early times, the primordial Universe consists in a very hot plasma where baryonic matter is ionised and strongly interacting with photons. Initial perturbations seeded by inflation propagate in

[17] http://www.darkenergysurvey.org/
[18] http://www.euclid-ec.org/
[19] https://www.cfa.harvard.edu/supernova/public.html
[20] http://www-supernova.lbl.gov/

the fluid following sound-waves caused by the pressure gradients. Eventually, the Universe becomes neutral at recombination and baryonic matter and photons stop interacting shortly after: at decoupling. At this moment, oscillations also freeze and leave imprinted the scale of the sound horizon $\chi_{\rm BAO}$ in the distribution of matter [25, 26]:

$$\chi_{\rm BAO} = \frac{c}{\sqrt{3}} \int_0^{a_{dec}} \frac{da}{a^2 H(a) \sqrt{1 + 3\Omega_b/(4\Omega_\gamma)}} \tag{4}$$

being $a_{\rm dec}$ the scale factor at decoupling.

This scale can be found at different cosmological times in the distribution of galaxies (and other tracers of matter) as a bump in the correlation function at large scales. We can use it as a standard ruler to determine the angular distance-redshift relation

$$d_A(z) = \frac{1}{1+z} \int_0^z \frac{c\, dz'}{H(z')} \tag{5}$$

being related to Equation 3 through $d_M(z) = d_A(z) \cdot (1+z) = d_L(z)/(1+z)$. Both BAO and SNIa measurements are shown together in Figure 2.

Even though disentangling the highly non-linear physics involved in galaxy and structure formation might be arduous, BAO measurements from the Large Scale Structure (LSS) became soon a very powerful tool to constrain Cosmology. This is partially due to the size of $\chi_{\rm BAO}$ relying on more basic principles. BAO was first detected in the galaxy distribution by the 2dFGRS collaboration[21] [27] and SDSS[22] [28], later more precise measurements were performed by 6dFGS[23] [29], WiggleZ[24] [30] and BOSS[25] [31]. Additionally, BOSS measured BAO from Lyman-α forests: 3D reconstruction of intergalactic blobs of neutral hydrogen that imprint absorption lines in spectra from distant quasars [32, 33].

[21] http://www.2dfgrs.net/
[22] http://www.sdss.org/
[23] http://www.6dfgs.net/
[24] http://wigglez.swin.edu.au/site/
[25] http://Cosmology.lbl.gov/BOSS/

Cosmic Microwave Background (CMB). After decoupling, light travels free leaving a bath of photons in the Universe that we see now at the temperature of $T_{\rm CMB} = 2.73K$: the CMB. This was one of the predictions of the Hot Big Bang theory confirmed experimentally by the discovery of Penzias and Wilson [34].

This temperature is isotropic to a fraction of $\sim 10^{-5}$, but there are small fluctuations originated during inflation and later propagated in the matter-photon plasma as already explained. After decoupling, light-rays are slightly bent by the Large Scale Structure (CMB lensing [35]) and photon energy slightly modified by structure collapse (ISW effect, [36]). Today, we can measure the CMB fluctuations containing information about those three epochs. The most visible effect is precisely the BAO peak. However, in the rest of this thesis, unless otherwise stated, BAO will refer to the imprint left in the matter distribution.

Amongst other CMB experiments, we highlight the satellites COBE[26] [37] that first detected the CMB anisotropies, WMAP[27] [38] that measured the first three peaks of the power spectrum and Planck that completed the power spectrum up to $\sim 0.07^\circ$ scales [11].

Towards Precision Cosmology with Large Scale Structure

Currently, the strongest constraints on Cosmology are set by CMB. This is in part due to the fact that CMB physics can be easily understood and modelled from linear perturbation theory, this boosted these type of experiments, pioneering Precision Cosmology. But its information exploitation reached a maximum and now attention focuses in higher order and more subtle effects (polarization, spectral distortions, etc.)

On the other hand, our understanding of the Large Scale Structure (LSS) is notably improving with time. This is due to a better control of the systematic experimental errors, more precise and specialised instruments, but also due to a better understanding of the astrophysics involved and a more precise knowledge of the Cosmology behind.

[26] http://lambda.gsfc.nasa.gov/product/cobe/
[27] http://map.gsfc.nasa.gov/

Moreover, whereas CMB provides a 2-dimensional map with information mostly about physics before and during decoupling, LSS explores the full 3-dimensional space at later times, where the influence of Dark Energy appears. LSS can be explored via different tracers: galaxies, quasars, galaxy clusters, Ly-α forests, HI-emission gas, etc.

We have already seen the power of galaxy surveys to measure BAO. But, with a better understanding of the LSS and the appropriate data, galaxy surveys can set more stringent constraints. When it comes to determining the nature of Dark Energy and to distinguishing between a Cosmological Constant and modified gravity, all the information is useful, since modifications of gravity can show their signature in many different measurements.

Amongst the targets in the coming galaxy surveys we find: measuring the full power spectrum of matter to constrain n_s, making 3D tomography of the Universe with weak lensing, measuring the 3-point correlation function in search of non-gaussianities, determining growth factors to constrain modified gravity, estimating the halo mass function (especially the high end), etc. [19, 20]

But for Precision Cosmology from LSS observations we need a counterpart from theory. Although, there are analytical models of structure formation [39], their predictions are limited, and we need to rely on N-Body simulations and other associated computing tools (see Section 1.1 and references therein). In Figure 3 we find a representation of data from galaxy surveys (blue) and its simulated counterpart (red), finding similarities in the distribution of galaxies, which follow complex and highly non-linear structures.

The tools that we use for simulations need to be validated in order to verify if they are appropriate for Precision Cosmology. Along these lines, in Chapter 1 we test the performance of different Halo Finders and Merger Tree builders used by the community. Halo Finders are tools used to identify bound/dense objects within the dark matter distribution of a simulation, that serve as hosts for galaxies. Merger tree builders are tools designed to follow the history of those objects along the cosmological time in the simulation. From the same underlying dark matter field we apply different combination of these techniques and analyse their outcome Merger Trees

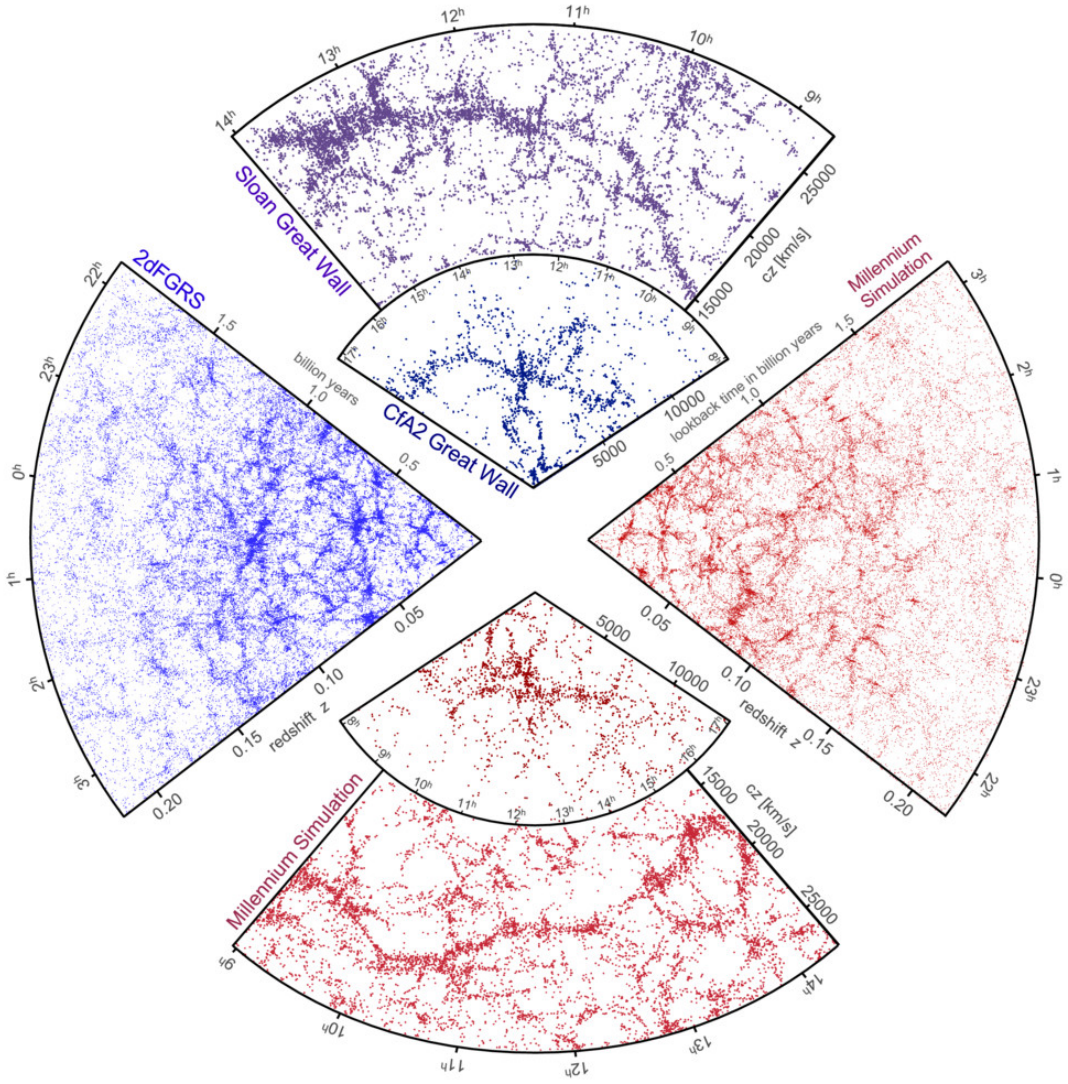

Figure 3: Representation of catalogues of galaxies from galaxy surveys in blue (SDSS and 2dFGRS) and the millennium N-Body simulation in red (`http://wwwmpa.mpa-garching.mpg.de/millennium/` [44]) Radial coordinates represents redshift z or, equivalently, recession Hubble velocity or cosmological look-back time. Angular coordinates, represents angle of observation. Both in simulations and observations we find a very complex and non-linear distribution of galaxies forming superclusters, filaments, walls and voids.

(scheme representing the history of halos). These are analysed by means of the halo age, merger rate and mass evolution.

Additionally, measurements from LSS need error bars and covariance matrices accounting for systematic errors, cosmic variance and their interplay. In order to estimate them we do not only need one simulation but hundreds or even thousands. Precise N-Body simulations are very costly in terms of computing resources, and running that many is prohibitive (see Section 2.1 and references therein). Hence, we need a new generation of simulating tools to generate approximate synthetic catalogues. In Chapter 2 I present HALOGEN, a technique designed to generate halo catalogues with the correct 2-point correlation function at large scales, reducing the CPU-hours required by a factor of $\sim 10^{3-5}$ compared to an N-Body simulation, and memory in a factor $\sim 10^{1-2}$ (Table 2.5). Other approximate methods for fast generation of halo mock catalogues from the literature are also presented and compared in Section 2.6.

Finally, in Chapter 3, I present the application of HALOGEN in the context of the Dark Energy Survey data analysis. Firstly, the catalogues are adapted with three additional observational features: construction of a lightcone, simulation of photometric redshift and the implementation of a Halo Occupation Distribution scheme (HOD, see Section 3.2.3 and references therein) fitted to reproduce the observed galaxy clustering. Then, a batch of mock catalogues is generated and we show its applicability to: gain insight into the modelling, optimise the analysis methodology and compute error bars and covariance matrices for the Large Scale Structure analysis.

Chapter 1

Merger Trees and Halo Finder Comparison

1.1 Introduction: Cosmological Simulations

In the early stages of evolution of the Universe, the homogeneous and isotropic assumption represents a good approximation. For studies of CMB where perturbations are very small ($\delta \sim 10^{-5}$), we can use linear perturbation theory to model the distribution of matter in the Universe. However, these small fluctuations continue to grow due to gravitational collapse forming the complex cosmic web (with filaments, knots and walls) that we find around us at the present epoch (Figure 3).

The details of structure formation can not be properly modelled from perturbation theory because collapsed objects enter in the highly non-linear regime of gravity. In this regime, we can only rely on N-Body simulations, where particles are let evolved with gravity step by step. We will shortly review the methods to perform N-Body simulations in Section 1.1.1

The final outcome of an N-Body simulation represents the dark matter density field as shown in Figure 1.1, which is not a direct observable. Hence, for each type of N-Body simulation and associated observation, a post-processing is needed (Section 1.1.2). Halo finders and merger tree builders are part of the analysis pipeline of

N-Body simulations. The former finds collapsed objects called *halos* within the dark matter distribution at a given time-step or snapshot (Section 1.2). The latter links those halos across different time-steps and identifies the merger of halos generating a scheme called *merger tree* (Figure 1.2 and Section 1.3).

There is a wide variety of methods used by the community for halo finding and merger tree building. In this chapter we analyse the differences and similarities in the outcome merger trees for the different combination of methods, see Section 1.1.3 for a more detailed description of the context and motivation of this study. This analysis is done on one hand from the geometrical point of view (Section 1.4) and on the other hand studying the halo mass evolution Section 1.5. Finally, conclusions are presented in Section 1.6.

1.1.1 N-Body Simulations

N-Body simulations are used in many fields of Astrophysics and other fields of physics. Depending on the area, there are different physical processes that may be relevant, and the simulations will have different requirements. For Large Scale Structure, we only simulate *dark matter* particles, for which only gravity and the expansion of the universe are relevant. These particles are not fundamental particles, but collisionless tracers of the phase-space, with very high masses exceeding a million (and often a billion) solar masses. Even if baryons –which represent a relevant fraction of the matter of the Universe– are not collisionless, in these simulations the hydrodynamics of baryons is neglected since its effects are only relevant at small scales ($\lesssim 2Mpc$) and represent a severe increment in the computing time. See [45] for a thorough study of N-Body simulations in different fields and a detailed derivation of the computations presented below. A review more specialised in N-Body methods in Cosmology is [46], and [47] is a more recent review more contextualised with experiments.

Cosmological simulations represent the Universe in a box of constant comoving volume sampled with N particles. In order to model an infinite and boundless Universe, we impose periodic conditions, i.e. a particle leaving one side of the box appears in the opposite side, and the gravitational potential is generated not only by the

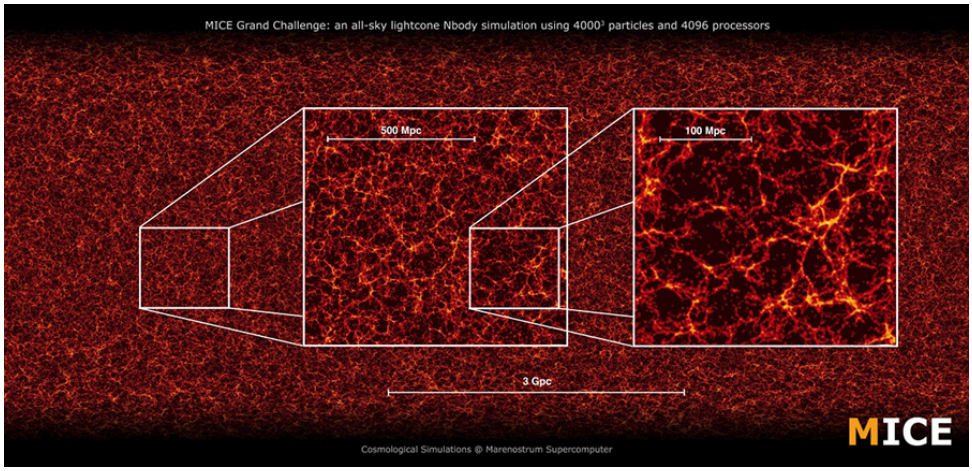

Figure 1.1: MICE Grand Challenge N-Body Simulation Dark Matter distribution, brighter parts represent denser regions (Table 2.1). This figure explores the very large scales up to $3072 Mpc/h$, which is the simulation size (being the extension replicas following the periodical conditions, see text), as well as intermediate scales ($100 Mpc/h$), where the deviations from homogeneities are more pronounced.

particles in the box, but also by an infinite number of replicas of the same box. Each particle is represented by its comoving coordinate $\vec{x} = \vec{r}/a$ and comoving velocity $\vec{u}$, being $\vec{r}$ its physical position. In this comoving frame, the equations of motion are left as

$$\begin{aligned} \frac{d\vec{x}}{dt} &= \vec{u} \\ \frac{d\vec{u}}{dt} &= -2 \cdot H \cdot \vec{u} - \frac{1}{a^3} \nabla_x \, \phi \end{aligned} \tag{1.1}$$

where $\phi(\vec{x})$ is the Newtonian potential determined by the density perturbations respect to the mean $\bar{\rho}$:

$$\nabla_x^{\,2} \phi = 4\pi G(\rho(\vec{x}) - \bar{\rho}) \tag{1.2}$$

Having all the equations and physics that govern the system, we now need to specify the method to generate the initial conditions, numerically compute the potential and integrate the equation of motion.

Initial Conditions

The initial power spectrum of matter is well known as it is measured from the CMB. It can be easily calculated given that cosmology is known with codes as CAMB [48]. Starting from a completely uniform Universe, we use Lagrangian Perturbation Theory to perturb the field and generate a density field with the correct power spectrum.

Lagrangian Perturbation Theory (LPT) studies how particles (fluid elements) move across the fixed coordinate space, unlike Eulerian theory where the matter of study is the variation of density and velocity field at a given position [49, 50]. At $z = \infty$ particles are distributed on a regular grid with coordinates $\vec{q}$ (Lagrangian Position). As the Universe expands particles are displaced to their Eulerian position $\vec{x}$:

$$\vec{x}(t) = \vec{q} + \vec{\Psi}(t, \vec{q}) \tag{1.3}$$

being $\vec{\Psi}$ the so-called displacement field. Expanding it in a Taylor series:

$$\vec{\Psi} = \vec{\Psi}^{(1)} + \vec{\Psi}^{(2)} + ... \tag{1.4}$$

In Zel'dovich Approximation (ZA [51], traditional name given to 1^{st}-order LPT), we only keep the first term, whereas for 2^{nd}-order LPT (2LPT) we keep up to 2^{nd} order terms. Resolving the corresponding equations of motion, one arrives at:

$$\begin{aligned} \vec{\nabla}_q \vec{\Psi}^{(1)} &= -D_1(t)\delta(q) \\ \vec{\nabla}_q \vec{\Psi}^{(2)} &= \frac{1}{2} D_2(t) \sum_{i \neq j} (\Psi^{(1)}_{i,i} \Psi^{(1)}_{j,j} - \Psi^{(1)}_{j,i} \Psi^{(1)}_{i,j}) \end{aligned} \tag{1.5}$$

being D_1 and D_2 the 1^{st} and 2^{nd} order growth factors, respectively.

Initial conditions have to be generated at a high enough redshift such that perturbations are still in the linear regime. But if we set them at too large redshift the gravity solver will integrate numerical noise. The standard method for many years has been to use ZA at the redshift when density perturbations reach $\delta \sim 0.1$. However, it has been shown that transients may appear with that method [52], and using 2LPT is

becoming more standard nowadays.

Force computation

Once the initial conditions are set, particles move according to gravity via the force

$$\vec{F} = \frac{1}{a}\nabla\phi \tag{1.6}$$

There are different ways to compute it depending on the N-Body code. The most naive way to compute this force is using Newton's law for each particle i, under the so-called **Particle-Particle** (PP) approach:

$$\vec{F}(\vec{x_i}) = -\sum_{j\neq i} \frac{Gm_i m_j}{r_{ij}^2}\hat{r}_{ij} \tag{1.7}$$

being $\vec{r}_{ij} = \vec{x_i} - \vec{x_j}$ with $r_{ij} = |\vec{r}_{ij}|$ and $\hat{r}_{ij} = \vec{r}_{ij}/r_{ij}$

This method is very accurate, but very slow since it scales as $\mathcal{O}(N^2)$. **Tree** solvers [53–55] transform it to a $\mathcal{O}(N\log N)$ problem by arranging particles in a tree (a scheme where particles are hierarchically grouped by proximity) and only resolving groups of particles that subtend an angle $\theta > \theta_0$ from the position $\vec{x_i}$.

Another associated problem with both of these methods is that we have to manually add a softening to the force because we need to avoid strong accelerations caused by 2-body interactions at small distances (recall that we are simulating a collisionless fluid). This softening arises naturally in the **Particle-Mesh** (PM) approach [56], where Equation 1.2 is solved on a grid in Fourier space and the force derived from Equation 1.6. This method is really fast, but it lacks accuracy at small scales.

Combining the efficiency of PM and the accuracy of PP we find the **P^{3}M** method [57, 58]. It computes the large scale part of ϕ with the PM and the small scale contributions with Equation 1.7. This can be still computationally costly in very clustered regions and an another solution is the **TreePM** method [59, 60]. In this method, small scale forces are computed using the Tree approach, whereas large scales are computed with the Fourier transform as in PM. Nowadays, one of the

most commonly used N-Body codes is GADGET-2[1] which is a publicly available implementation of the TreePM algorithm. Different versions of this code have been used for the N-Body simulations presented in this thesis.

Another popular method is the **Adaptive Mesh Refinement** (**AMR**) method[61, 62], that is equivalent to a PM method, but increasing in real time the resolution of the mesh in the higher density regions. This is specially useful if we want to include hydrodynamical gas physics in an Eulerian approach.

Time integration

At every time-step after calculating the force, we need to move the particles according to the velocity and accelerate them according to their forces. It is actually found that it is desirably to do this alternatively using a *leap-frog* scheme. This is, position and momentum are not updated simultaneously, but with a delay of half the time-step Δt:

$$\begin{aligned} \vec{x}^{\,k+1/2} &= \vec{x}^{\,k} + \frac{\Delta t}{2}\frac{\vec{p}^{\,k}}{a^2 m} \\ \vec{p}^{\,k+1} &= \vec{p}^{\,k} + \Delta t \cdot \vec{F}^{\,k+1/2} \\ \vec{x}^{\,k+1} &= \vec{x}^{\,k+1/2} + \frac{\Delta t}{2}\frac{\vec{p}^{\,k+1}}{a^2 m} \end{aligned} \tag{1.8}$$

Note that the third part of the step $k = l$ its identical to the first part of the step $k = l + 1$, so they can be applied together forming the *leap-frog* scheme. It is in the second part where we need to compute the force at every step, as explained before.

1.1.2 Simulation Post-processing

The N-Body simulations give us the distribution of dark matter of the Universe. However, this is not a direct observable, and we need to include galaxies if want to compare them with observations. There are different methods to do so, which rely on different techniques that will be explained below:

[1] http://wwwmpa.mpa-garching.mpg.de/gadget/

- **Halo Finder**. Galaxies live in dark matter halos: self-bound, virialised and very dense objects with spheroidal shape. Halo finders are codes that identify these objects from the distribution of particles in the simulation at a given time-step or snapshot. Some halo finders also identify sub-halos (halos that lie in another halo). See Section 1.2 and [63] for a review.

- **Merger Tree** builders. A merger tree is a scheme that traces back a halo from the latest snapshot to the origin of all its progenitors, it tells us about the history of halos, including the age, merger rate, etc. A merger tree builder is a code that links halos across different snapshots. See Section 1.3 and [64] for a description of most methods.

- **Semi-analytical Methods (SAM)**. The aim of semi-analytical methods is to infer the properties of galaxies from the mass evolution and merger rate of the host halo as given by the merger trees. It is motivated from hierarchical structure formation and is a natural complement to Press-Schechter Theory [39]. This implies reducing very complex processes of gas physics (gas dynamics of rotation and merger, gas cooling, star formation, supernova feedback, etc.) to a few prescriptions and parameters. See [65] for a recent comparison with a summary of the different techniques.

- **Halo Occupation Distribution (HOD)**. The aim of HOD is mainly to place galaxies in a halo with the correct clustering. This model places galaxies in halos following a halo density profile (typically a NFW, [66]) with some free parameters that are fitted to data. It does not take into account any internal/temporal property of the halo, but only the final mass. Simple models are fitted to a specific observed sample, whereas other models more ambitious include magnitudes, colours, stellar mass, etc. [67–75]

- **Halo Abundance Matching (HAM)** or **Subhalo Abundance Matching (SHAM)**. This method relates a magnitude of the halos or subhalos (mass, v_{max}, etc) to a magnitude of the galaxies (luminosity, stellar mass, etc). This can be done either strictly by rank-order (e.g. the n^{th} most massive halo corresponds to the n^{th} most luminous galaxy) or stochastically by adding some scatter. Typically the scatter is used to match the clustering. [76–79]

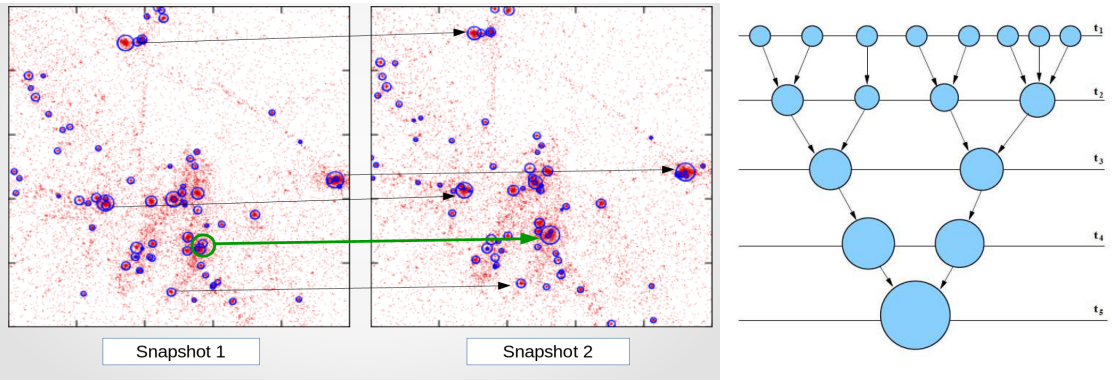

Figure 1.2: Merger Trees representation. On the left panel we see the particles of an N-Body Simulation as red dots, the halos as blue circles with radius R_{200c} (defined by Equation 1.9) and merger trees as arrows linking halos across snapshots (the green one represents a merger). On the right panel (from [80]) we find a chart representing a merger tree following the mergers of halos (circles) through different snapshots t_i.

1.1.3 The Comparison Project

We just saw in Section 1.1.2 that many techniques and prescriptions implemented in codes are used to analyse simulations. Different codes are used by different research groups in the community for the same purpose. Each of these codes make some approximations and assumptions but, do they lead to the same results? This is the question posed by the Mocking Astrophysics programme[2]. During a series of workshops and subsequent studies this programme has been analysing and validating the post-processing pipeline used by the community. Among the target of study we find the halo finders [81], merger tree builders [64] and Semi-analytical Models [65]. The analysis presented in the remainder of this chapter was done as part of this programme and as a consequence of the SussingMergerTree workshop[3]. The aim is to address the question of how the combination of halo finders and merger tree builders affect the properties of the final merger tree [1].

[2] http://www.nottingham.ac.uk/~ppzfrp/mockingastrophysics/
[3] http://popia.ft.uam.es/SussingMergerTrees

Sussing Merger Trees: the influence of the halo finder

The backbone of any semi-analytical model of galaxy formation is a merger tree of dark matter halos. Some modern semi-analytical codes [82–86] rely on purely analytical forms such as Press-Schechter [87] or Extended Press-Schechter [88] –see [89] for a comparison of such methods–, while other codes take as input halo merger trees derived from large numerical simulations (see [90, 91] for the historical origin of both approaches). Therefore, stable semi-analytic models require well-constructed and physically realistic merger trees: halos should not dramatically change in mass or size, or jump in physical location from one step to the next.

The properties of the merger trees built using a variety of different methods was addressed in [64]. While it was observed that different tree building algorithms produce distinct results, the influence of the underlying halo catalogue still remained unanswered. This is nevertheless an important question as different groups rely on their individual pipelines, which often includes their own simulation software, halo finding method and tree construction algorithm before the trees are fed to a semi-analytical model to obtain galaxy catalogues.

In a series of comparisons of (sub-)halo finders [e.g. 81, 92–95], which are all summarised in [96], it was shown that there can be substantial variations in the halo properties depending on the applied finder. This will certainly leave an imprint when using the catalogue to construct merger trees. As a fixed input halo catalogue was used for the first tree builder comparison, the question addressed here is to what extent merger trees are sensitive to the supplied halo catalogue.

In this work we include both steps of the tree building process, i.e. we will apply a set of different tree builders to a range of halo catalogues constructed using a variety of object finders. Please note that the underlying cosmological simulation remains identical in all instances studied here. We are investigating how much of the scatter in the resulting merger trees that form the input to semi-analytical models stems from the tree building code and how much stems from the halo finder. Or put differently, is a merger tree more affected by the choice of the code used to generate the tree or the code used to identify the dark matter halos in the simulation?

1.2 Halo Finding Techniques

As already explained, halo finders search dark matter halos within the particle distribution of a simulation snapshot. The exact definition of halos in simulations is actually set by the halo finder itself and can vary significantly from one finder to another, specially when it comes to subhalos (halos lying inside another halo). The subtleties of each code will be explained below, but we will introduce here the two basic types of halo finding techniques at the main halo level:

- Friends of Friend (FoF)[97]. It is a geometrical algorithm: particles separated less than $b \cdot D_{\text{mean}}$ are glued together to form halos. D_{mean} is the mean inter-particle separation, and b a free parameter typically set to 0.2.

- Spherical Overdensity (SO)[39]. This technique searches density peaks and defines halos as spherical objects with density higher than ρ_{ref} with a typical value of $200\rho_{\text{crit}}$.

While FoF can only give main halos, the SO method may also be used for subhalos. A minimum number of particles N_{min} must be considered for halos to be valid, generally $N_{min} = 20$ is chosen.

The halo catalogues used for this study are extracted from 62 snapshots of a cosmological dark-matter-only simulation undertaken using the GADGET-3 N-body code [98] with initial conditions drawn from the WMAP-7 cosmology [99]. We use 270^3 particles in a box of comoving width 62.5 h^{-1} Mpc/h, with a dark-matter particle mass of $m_p = 9.31 \times 10^8 h^{-1}\mathrm{M}_\odot$. We use 62 snapshots (000,...,061) evenly spaced in $\log a$ from redshift 50 to redshift 0.

While in previous comparison projects [e.g. 81, 93, 96] the same mass definition was imposed (or even used a common post-processing pipeline to assure this), it was not request any such thing this time, i.e. every halo finder was allowed to use its own mass definition.

On the one hand, AHF and ROCKSTAR define a *spherically truncated* mass through

$$M_{\rm ref}(< R_{\rm ref}) = \Delta_{\rm ref} \times \rho_{\rm ref} \times \frac{4\pi}{3} R^3_{\rm ref} \ , \tag{1.9}$$

adopting the values $\Delta_{\rm ref} = 200$ and $\rho_{\rm ref} = \rho_{\rm crit}$ (we will call this mass M_{200c}) and iteratively removing particles not bound to the structure. On the other hand, HBTHALO and SUBFIND return *arbitrarily shaped* self-bound objects based upon initial Friends-of-Friends (FoF) groups, assigning them the mass of *all* (i.e. no spherical truncation) particles gravitationally bound to the halo.

Furthermore, some halo finders include the mass of any bound substructures in the main halo mass whereas others do not include the mass of any bound substructures. Technically, finders for which particles can only belong to one halo are termed *exclusive* while finders for which particles can belong to more than one halo are termed *inclusive*. As substructures can typically account for 10% of the halo mass this choice alone can make a substantial difference to the halo mass function.

Given these definitions we can now describe the general properties of the halo finders applied to the data:

- AHF [100, 101] is a configuration-space Spherical Overdensity (SO) adaptive mesh finder. It returns inclusive gravitationally bound halos and subhalos spherically truncated at R_{200c} (thus, the mass returned is M_{200c}).

- HBTHALO [102] is a tracking algorithm working in the time domain that follows structures from one time-step to the next. It returns exclusive arbitrarily shaped gravitationally bound objects. It uses FoF groups for the initial particle collection.

- ROCKSTAR [103] is a phase-space halo finder. A peculiarity of this code is that –unlike AHF, HBTHALO and SUBFIND– the mass returned for a halo does not correspond to the sum of the mass of the particles listed as belonging to it. While it uses the same mass definition as AHF (inclusive bound M_{200c} mass), the particle membership list of the halo is exclusive and is made up by proximity of particles in phase-space to the halo centre.

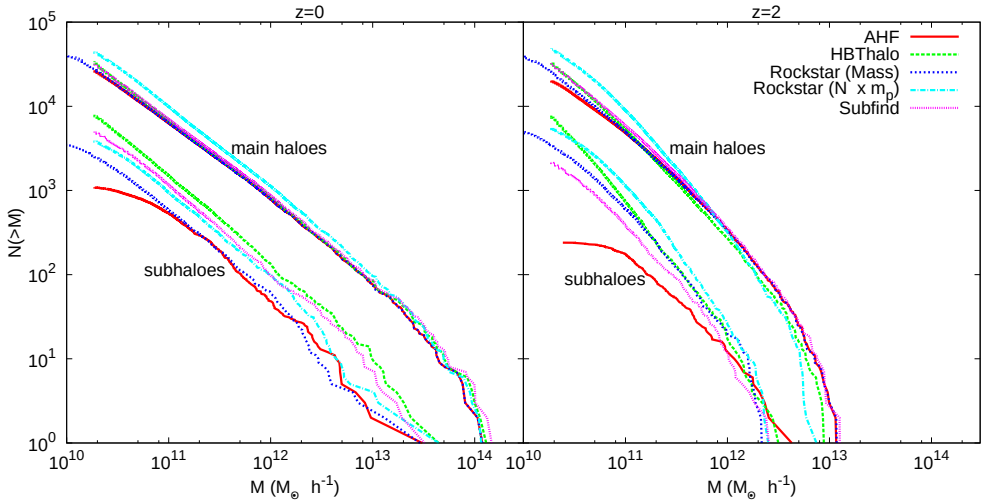

Figure 1.3: Cumulative mass functions at redshift $z = 0$ (left panel) and $z = 2$ (right panel) for the four halo finders. There are two lines for ROCKSTAR corresponding to the two mass definitions discussed in the text: one corresponding to M_{200c} (*Mass*) and one based upon the particle list ($N \times m_p$, being N the number of particles and m_p the particle mass). The upper set of curves in each panel is based upon main halos whereas the lower set of curves in each panel refers only to subhalos.

- SUBFIND [104] is a configuration-space finder using FoF groups as a starting point which are subsequently searched for subhalos. It returns arbitrarily shaped exclusive self-bound main halos, and arbitrarily shaped self-bound subhalos that are truncated at the isodensity contour that is defined by the density saddle point between the subhalo and the main halo.

To give an impression of the differences in the halo catalogues, we present in Figure 1.3 the cumulative mass function for the four halo finders at redshift $z = 0$ (left panel) and $z = 2$ (right panel); further, we separate subhalos from main halos and present their cumulative mass spectrum in the upper and lower set of curves of each panel, respectively. A threshold of 20 particles has been set (equivalent to $M = 20m_p = 1.86 \times 10^{10} h^{-1} M_\odot$) for halos to be considered. In order to highlight the peculiarity of ROCKSTAR (for which the returned mass does not correspond to the sum of the mass of the particle membership) we included two lines for ROCKSTAR: one based upon summing individual particle masses (cyan dash-dotted) and one with the mass M_{200c} as returned by ROCKSTAR (blue dotted, extending to masses below

the 20 particle threshold). Given that some tree builders only use particle membership information for a halo whereas others combine this with a table of global properties (including halo mass), this choice of mass definition will also contribute to the differences in the final trees.

We find that other than for the largest 100 main halos the different mass definitions make little difference for the main halos at $z = 0$ unless the mass taken from the returned ROCKSTAR particle membership is used. This mass is systematically higher that the other estimates (and ROCKSTAR's own returned mass). The differences in mass for main halos are slightly more pronounced at $z = 2$.

For subhalos there are noticeably different mass functions: AHF is incomplete at the low-mass end, with a trend that appears to worsen as the redshift increases[4]. However, despite generally finding more subhalos the other finders do not appear to have converged to a common set. Part of this relates to the rather ambiguous definition of subhalo mass: whereas for main halos it simply appears to be a matter of choice for $\Delta_{\rm ref}$ and $\rho_{\rm ref}$ (or some other well-defined criterion for virialisation/boundness/linkage), subhalos – due to the embedding within the inhomogeneous background of the host – cannot easily follow any such rule. Again, each finder has been allowed to pick its favourite definition for subhalo mass. But please note that the variations seen here are not the prime focus of this study; they should nevertheless be taken into account when interpreting the results presented and discussed below. Further, the scatter in subhalo mass functions seen in previous comparisons was much reduced due to the use of a common post-processing pipeline that ensured a unique subhalo mass definition [93, 94, 96].

All these differences should and will certainly leave an imprint and be reflected in the outcome when building merger trees.

[4]It was checked that a more restrictive parameter set for AHF leads to the recovery of the missing low-mass subhalos at high redshift. As already shown by [101] (Fig.5 in there) there is a direct dependence of the applied refinement threshold used by AHF to construct its mesh hierarchy (upon which halos are based) to the number of low-mass objects found.

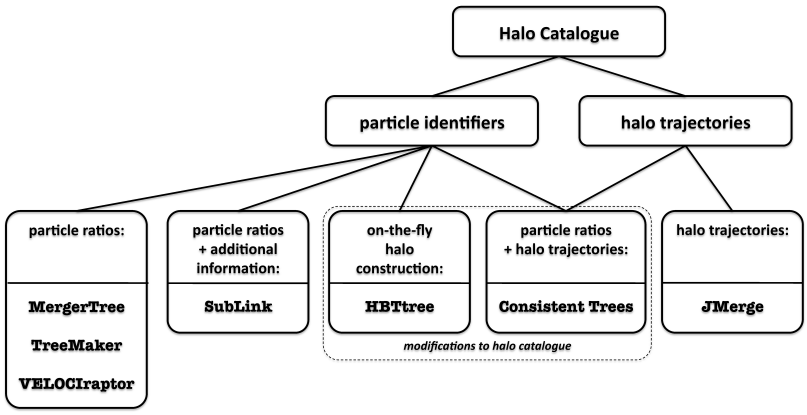

Figure 1.4: A summary of the main features and requirements of the different merger tree algorithms. For details see the individual descriptions in the text.

1.3 Merger Tree Builders

In the first merger tree comparison paper [64], we can find an extensive description of most merger tree builders available and a terminology convention to describe them, that we also use here. A lot of the methodology is similar across the various codes used for this study, the main features and requirements have been captured in Figure 1.4. We first categorise tree builders into either using halo trajectories (JMerge, and Consistent Trees) or individual particle identifiers (together with possibly some additional information; all remaining tree builders). Consistent Trees is the only method that utilises both types of approach. HBT constructs halo catalogues and merger trees at the same time as it is a tracking finder that follows structures in time. A cautionary note regarding HBT: it can be applied both as a halo finder or a tree builder and includes elements of both so we will always specify whether we refer to one or the other by appending 'halo' or 'tree', as necessary.

The codes themselves are best portrayed as follows:

- Consistent Trees forms part of the Rockstar package. It gravitationally evolves positions and velocities of halos between time-steps, making use of information from surrounding snapshots to correct missing or extraneous halos in individual snapshots [105].

- HBTtree is built into the halo finder HBT. It identifies and tracks objects at the same time using particle membership information to follow objects between output times.

- JMerge only uses halo positions and velocities to construct connections between snapshots, i.e. halos are moved backwards/forward in time to identify matches that comply with a pre-selected thresholds for mass and position changes.

- MergerTree forms part of the AHF package and cross-correlates particle IDs between snapshots.

- SubLink tracks particle IDs in a weighted fashion, giving priority to the innermost parts of subhalos and allowing branches to skip one snapshot if an object disappears.

- TreeMaker consists of cross-comparing (sub)halos from two consecutive output times by tracing their exclusive sets of particles.

- VELOCIraptor is part of the VELOCIraptor/STF package and cross-correlates particle IDs from two or more structure catalogues.

Two codes were allowed to modify the original catalogue: Consistent Trees and HBTtree. Consistent Trees adds halos when it considers they are missing: i.e., the halo was found both at an earlier and at a later snapshot. Consistent Trees also removes halos when it considers them to be numerical fluctuations: i.e., the halo does not have a descendant and both merger and tidal annihilation are unlikely due to the distance to other halos. HBTtree for *external* halo finders (i.e. halo catalogues not generated by its own inbuilt routine) takes the main halo catalogue and reconstructs the substructure. This produces an exclusive halo catalogue in which the properties of the main halos may also have changed.

1.4 Geometry of trees

In this section we present the geometry and structure of merger trees and the resulting evolution of dark matter halos. This includes the length of the tree (Section 1.4.1) and the tree branching ratio (Section 1.4.2). Further, it is shown graphically how halo finders and tree builders work differently, to illustrate the features found in the comparison.

1.4.1 Length of main branches

One of the conceptually simplest properties of a tree is the length of the main branch. It measures how far back a halo can be traced in time – starting in this case at $z = 0$. This property not only relies on the performance of the halo finder and its ability to identify halos throughout cosmic history, but also on the tree builder correctly matching the same halo between snapshots. [64] found that the different tree building methods produced a variety of main branch lengths, ascribing some of the features to halo finder flaws. We shall verify this now.

Figure 1.5 shows a histogram of the main branch length l, defined as the number of snapshots a halo main branch extends backwards in time from snapshot 61 ($z = 0$) to snapshot $61 - l$. This is roughly equivalent to an age, given that the last 50 snapshots are separated uniformly in expansion factor, $a = 1/(1 + z)$. On the left, we selected the 1000 most massive main halos, whereas on the right we see the results for the 200 most massive subhalos. The main halo population coincides from one halo catalogue to another in at least 85% of the objects. The subhalo population is more complicated and, in some cases, they only agree in 15% of the objects from one finder to another. However, if we focus on comparing AHF with ROCKSTAR or HBTHALO with SUBFIND, we find a better agreement between catalogues, rising to $\sim$ 95% for main halos and $\sim$ 70% for subhalos. Due to these differences, the applied number threshold translates to mass thresholds M_{th} that are different from finder to finder (see also Figure 1.3); we therefore list the corresponding values in Table 1.1. Furthermore, when using HBTTREE, the individual masses of the halos can change and so does the mass threshold. In what follows we will consistently use

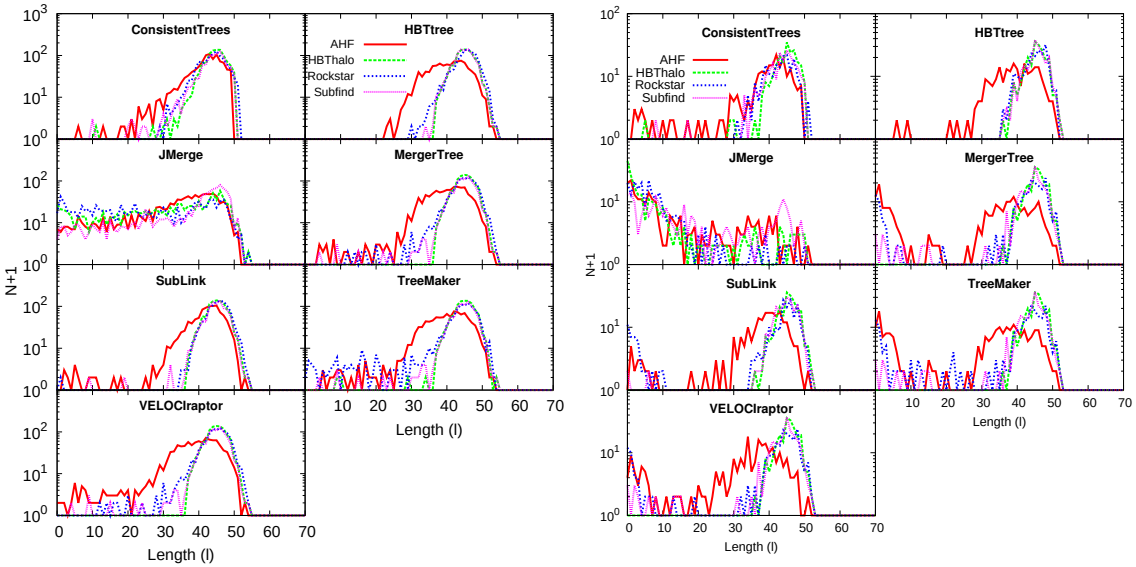

Figure 1.5: Histogram of the length of the main branch. The length l is defined as the number of snapshots a halo can be traced back through from $z = 0$. The left group of panels show the 1000 most massive main halos. The right group of panels show the 200 most massive subhalos. These number selections are equivalent to the mass cuts shown in Table 1.1. Different panels contain results from different tree building methods (as indicated), while within each panel there is one line for each halo finder (as marked in the legend).

	AHF	HBTHALO	ROCKSTAR	SUBFIND
M_{th}^{main}	7.93	8.25	7.90	9.61
$M_{th,\text{HBT}}^{\text{main}}$	7.52	8.25	10.64	8.30
M_{th}^{sub}	3.09	6.91	3.00	5.30
$M_{th,\text{HBT}}^{\text{sub}}$	2.75	6.91	2.68	5.90

Table 1.1: Mass threshold in units of $10^{11}h^{-1}M_{\odot}$ needed to select at $z = 0$ the 1000 most massive main halos (rows 1 and 2) and the 200 most massive subhalos (rows 3 and 4) for different halo finders (columns). Odd rows show the threshold for a general tree builder, whereas even rows show the threshold for HBTTREE

these mass thresholds, even at higher redshift.

As expected by the hierarchical structure formation scenario induced by cold dark matter, most large mass objects can be traced back to high redshift. This is not surprising and was already reported in [64], but here we can appreciate that this result depends on the choice of the halo finder and we will elaborate on this below.

As a general observation, for both main halos and subhalos, it is apparent that HBTHALO leads to the best results: nearly all massive halos are found and followed from an early origin. We attribute this to the fact that by its very nature as a tracking finder HBTHALO is designed with the intention of building a merger tree in mind. SUBFIND tends to give similar results but with occasional early truncation. These truncations become more pronounced for AHF and ROCKSTAR. Further, AHF tends to terminate each tree slightly earlier, even if it was well followed back in time, because of the incompleteness at low mass end (Figure 1.3). For AHF missing low-mass objects at high redshift cannot be the small progenitors of the high-mass low-redshift objects followed in Figure 1.5.

Differences between subhalos and main halos are also apparent. First, the subhalo curves in general appear more noisy, in part due to having fewer objects, but also because they are always placed in a more complicated environment which enhances the stochasticity. The difficulty in following subhalos then causes more cases with low l, especially for AHF and ROCKSTAR. One could naively think his excess of low-l subhalos for AHF and ROCKSTAR could be the result of a much smaller M_{th}^{sub} threshold (see Table 1.1). However, it was verified that using the same threshold for all catalogues, only mitigates that difference without completely erasing it.

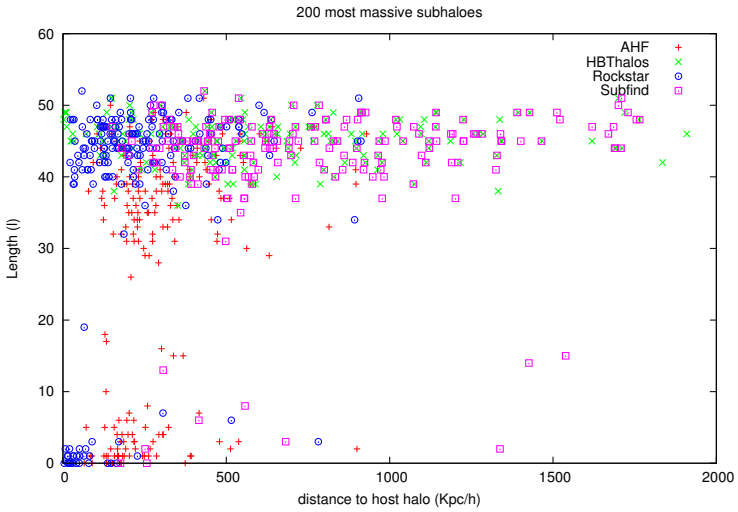

Figure 1.6: Distance to the centre of the host halo vs. length of the tree l for the 200 most massive subhalos. We show the results for the four halo finders (see legend) for the MergerTree builder.

Subhalo finding becomes especially difficult as the subhalo approaches the centre of the host halo, as has been shown in Fig.4 of [106] and Fig.7 of [93]. In particular, SUBFIND underestimates the mass of subhalos close to the centre of their host halo. Given that the 200 most massive subhalos are not the same for all finders, the subhalos selected for SUBFIND tend to be further from the host halo centres (see Figure 1.6), and therefore they are easier to trace. AHF and especially Rockstar find many (massive) subhalos near the centre but, due to the difficulties in that region, a fraction of them cannot be provided with a credible progenitor in an earlier snapshot, resulting in early tree termination. Finally, the HBThalo selection is composed of subhalos at short, medium and large distances from the host halo centre but, by construction, they are always required to be traceable.

On the tree builder side, JMerge allows halos to only shrink their mass by a factor of up to 0.7 and to grow by a factor of up to 4 in one snapshot, and it estimates their trajectories from global quantities (Section 1.3). This artificially truncates main branches too early for massive objects when it loses track of halos. This effect is enhanced for subhalos, whose trajectories are difficult to estimate due to the non-linear environment and the fact that their mass is more likely to grow or shrink

abruptly (Section 1.5). Consistent Trees and HBTtree essentially eliminate the low-l cases for nearly all the halos (and subhalos). This is due to their freedom to modify the catalogue in such a way as to avoid exactly these occurrences.

In order to better illustrate the factors that influence the main branch length, l, we present in Figure 1.7 a graphical representation of the performance of the various halo finders and tree builders. This slice shows two halos of similar size passing through each other in the process of a merger (the same merger as shown in Fig.4 of [64]). These two halos are identified at $z = 0$ with a blue line and a red line, and then traced back by the merger tree. For this example MergerTree, TreeMaker and VELOCIraptor gave identical results and so we only show the MergerTree result.

We find a wide variety of situations: in some cases every halo is correctly traced (e.g. Consistent Trees with AHF) but in others the tracing fails (e.g. JMerge with AHF). In the success or failure of the tracing the influence of both the halo finder and the tree builder are important:

- AHF considers one of the merging halos to be the main halo (blue) and the other to be a subhalo (red). In snapshot 060 the subhalo found is quite small, so that most of the tree building codes do not link it with the (much larger) halo in the next snapshot (061). In simple codes (JMerge, MergerTree...) this leads to an artificial truncation of the tree. Consistent Trees artificially adds one halo to snapshot 060 to replace the small subhalo whereas SubLink jumps snapshot 060 for this object. In this way both codes continue the tree. HBTtree recomputes the substructure, creating a more traceable subhalo.

- HBThalo is able to identify at snapshot 060 two big and well defined halos of almost the same size (only possible for exclusive halo catalogues). This is due to the tracking nature of the finder and ensures the correct follow-up by most tree builders. Only JMerge encounters problems due to the non-smooth trajectories of the halos.

- Rockstar uses phase-space information so that even when the halos are overlapping (snapshot 060) it is able to distinguish them by their velocities. This

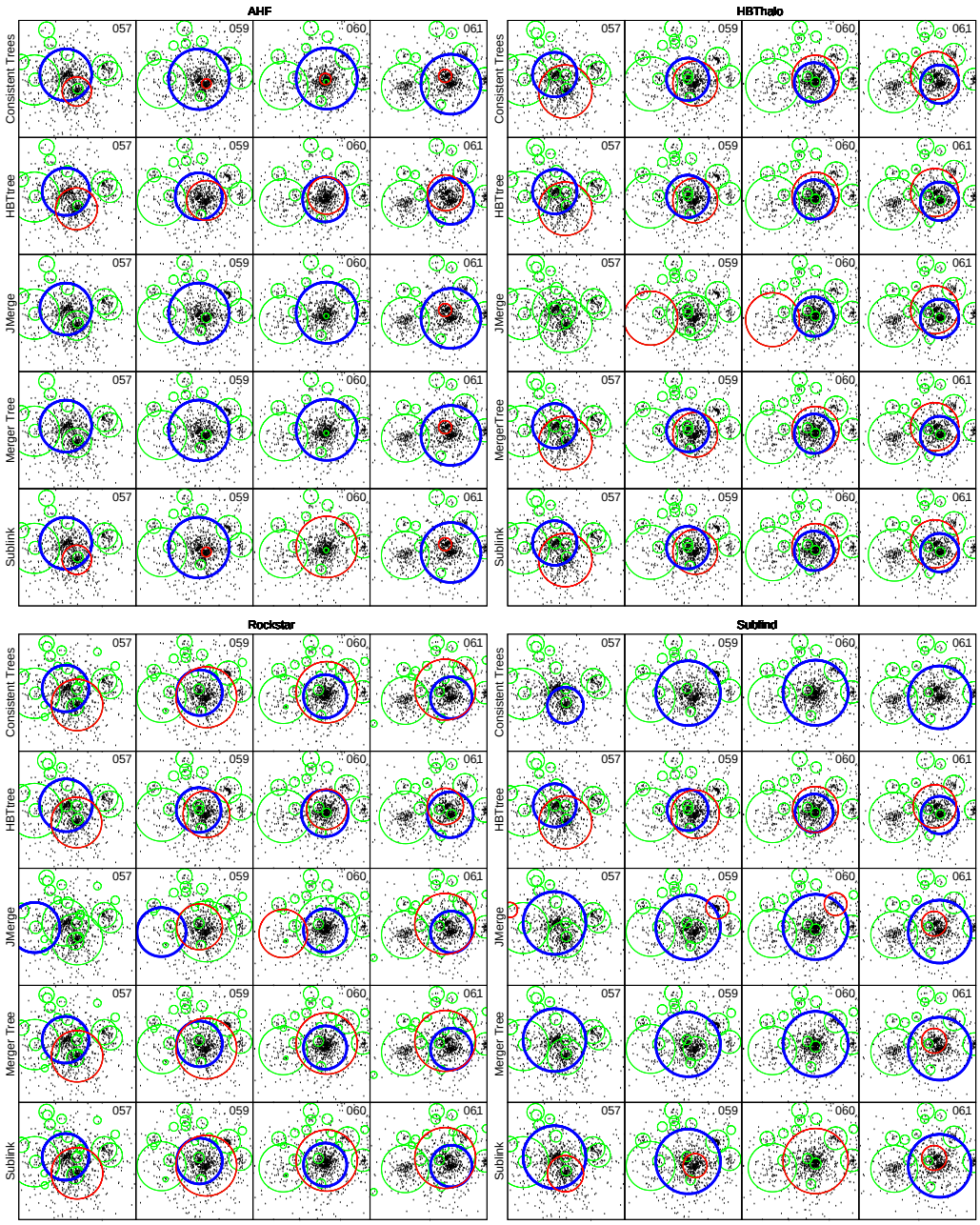

Figure 1.7: Projected image of a 1.2 Mpc/h-side cube from the N-Body simulation. Halos are represented by circles of radius corresponding to R_{200c}. This is an example of a merger between two halos that are found at $z = 0$ (snapshot 061) and linked across snapshots by the tree builders: the blue and red colours represent the two trees. Other halos found are represented in green. Each subfigure presents a single halo finder, with each row representing the indicated tree builder. In each row time evolves from left to right, with each cell a different snapshot.

allows almost all tree codes (besides JMerge) to follow the evolution of the halos.

- SUBFIND gives similar problems to AHF: the subhalo at snapshot 060 is too small to be considered a credible progenitor. For this catalogue, Consistent Trees is not able to deal with it and completely removes the red tree. HBTtree patches over that problem the usual way while JMerge associates the halo to a progenitor incorrectly and MergerTree truncates the tree. SubLink, by omitting snapshot 060, is able to follow the history correctly.

This example neatly illustrates the difficulties that arise when dealing with subhalos. However, the left panel of Figure 1.5 tells us that there are also situations in which the main halo branch is truncated. We studied several of these cases and found two main types: in the first type the main halo lies in the vicinity of a bigger halo, and is likely to have entered it and become a subhalo a few snapshots before. In this case the problems encountered are similar to those illustrated in the subhalo example above, but here the in-falling halo has been classified as a main halo at $z = 0$. The other type occurs when at some point the halo was wrongly associated to some other smaller halo as happened with the red halo in Figure 1.5 for the combination JMerge-HBThalo. In this case the incorrect halo assignment never gets corrected and typically the much smaller halo has a much shorter prior history.

Already at this stage of the analysis we can draw some conclusions from this subsection:

- In general, the influence of the halo finder is at least as (if not more) important than the tree building algorithm.
- Main halos are easier to trace.
- The way the halo finder deals with substructure is crucial for merger trees.
- Tree building *tricks* such as the creation of artificial halos or omitting snapshots help in some cases, but are not infallible.

- AHF and ROCKSTAR catalogues lead to earlier tree truncation for most tree builders. This is especially true for subhalos, because they try to find subhalos close to the host halo centre and are not able to provide them with credible progenitors.
- SUBFIND tends to find more subhalos in the outer regions of the host, which are easier to track.
- HBT appears to be very well designed to not truncate a tree too early, both as a halo finder and as a tree builder (as seen in Figure 1.5 and 1.7).
- CONSISTENT TREES also stands out in avoiding low-l cases (Figure 1.5).
- JMERGE faces problems in complex environments.

1.4.2 Branching ratio

Another simple tree property, which is nevertheless very important for characterising the structure or geometry of a tree, is the number of direct progenitors N_{dprog} (or local branches) that a halo typically has. Figure 1.8 shows the normalised (divided by the total number of events) histogram of N_{dprog} for all halos in the range $0 \leq z \leq 2$. For all the various combinations of tree building method and halo finder the most common situation is to have just one single progenitor, corresponding to a halo having no mergers on this step (which can happen multiple times during a halo lifetime). The second most common situation is for a halo to have no progenitors, which corresponds to a halo passing above the detection threshold and appearing for the first time, which can happen only once. As for other properties studied in this study, our results would certainly change if we were to use a different set of output times, so the importance does not lie in the individual tree results, but in their differences. For an elaborate study of the optimal choice for the temporal spacing of snapshots to construct merger trees see [107] or [86].

It is noticeable that the ROCKSTAR catalogue (blue dotted line) yields a tree with significantly large branching ratio for the tree builders SUBLINK, TREEMAKER, and VELOCIRAPTOR. Also, besides using a very similar technique, MERGERTREE

shows a more moderate branching ratio. By removing objects with mass lower than $20m_p$ (cyan dash-dotted line), we verified that this high branching ratio is caused by objects with very low mass as these high-$N_{\rm dprog}$ cases disappear. Recall that, even though all the halo finders cut their catalogue at 20 particles, for ROCKSTAR the mass M_{200c} can be lower if some of those particles lay outside R_{200c}. This small change, in general, moves the curves for ROCKSTAR from the highest branching ratio to the lowest one. Note that the mass limited tree shown in cyan is not equivalent to the other trees because the catalogue was reduced *after* running the tree building algorithm on it, hence giving non-self-consistent trees. Nevertheless, we do not expect great variations in Figure 1.8 between the cyan line and a fully self-consistent tree with the same mass limit. This serves as an illustration of the great influence of the lower mass limit, pointing out again the importance of the input halo catalogue in the resulting tree construction. To illustrate a high branching ratio case we have selected one of the extreme cases with $N_{\rm dprog} > 30$ in Figure 1.9. It corresponds to one of the two most massive halos (depending on the halo finder) at snapshot 050 (z=0.32). Figure 1.9 shows all the direct progenitors of that halo and other halos found in the area. The blue halo is the main and most massive progenitor in the plot. The red and magenta circles represent other direct progenitors at snapshot 049 while green circles represent other (sub)halos detected in the same region. Magenta is used for halos whose mass is below $20m_p$ (only possible for ROCKSTAR), while red halos have larger mass. SUBLINK also has halos that were found at snapshot 048, but were not linked in snapshot 049, which were linked to the big halo at snapshot 050; these are marked as crosses.

Figure 1.9 tells us that, when comparing different halo catalogues, $N_{\rm dprog}$ tends to be correlated to the number of (small) halos available to be absorbed, i.e. the more green halos we find the more merging (red and magenta) halos we find. We further confirm that most secondary progenitors (red and magenta circles) are subhalos of the main progenitor (blue circle) and lie within R_{200c}. However, in some cases secondary progenitors were found outside the volume displayed (e.g. the halos missing in CONSISTENT TREES with AHF). But in general, the properties of these halos fit into the standard merging picture in which halos approaching a bigger one become satellites (subhalos), lose mass via tidal stripping and are eventually totally

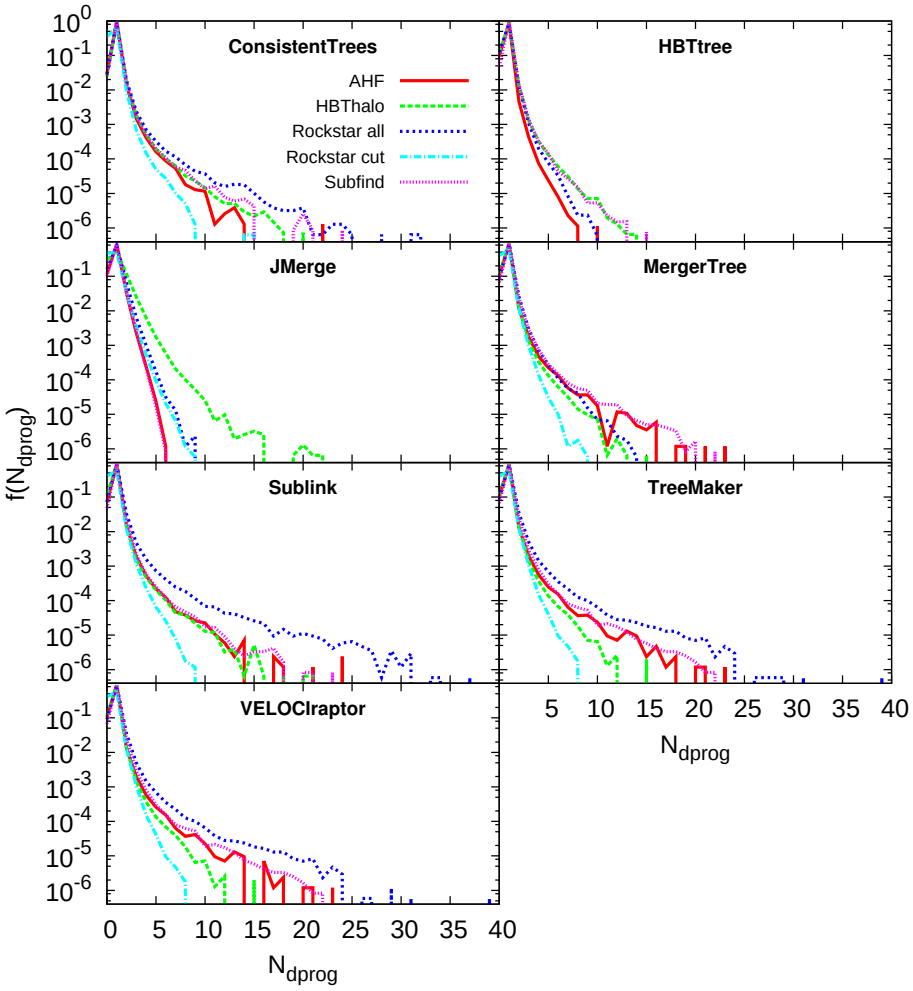

Figure 1.8: Normalised histograms of the number of direct progenitors $N_{\rm dprog}$ for all halos from $z = 0$ to $z = 2$ (snapshots from 061 to 031). Each panel corresponds to a single tree building method, within each panel each line represents a halo catalogue as indicated. For ROCKSTAR we show two lines, one with all the halos ('ROCKSTAR all') and one where halos with mass lower than 20 m_p were removed ('ROCKSTAR cut').

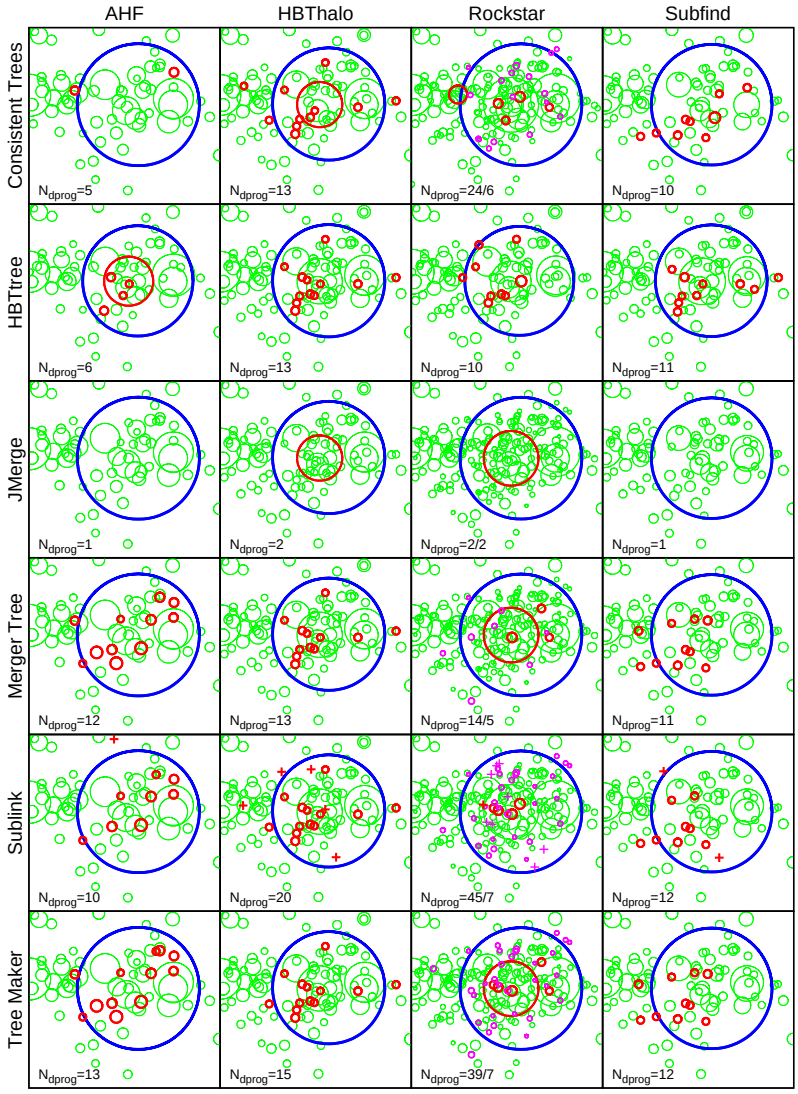

Figure 1.9: Projected image of a 3 Mpc/h-side cube from snapshot 049 centred on one of the most massive objects ($M > 10^{14}h^{-1}M_{\odot}$) for all the combinations of halo finder (column) and tree builder (row). Symbol and colour coding explained in the text. VELOCIRAPTOR (omitted) gives the same results as TREEMAKER. The label $N_{\rm dprog}$ indicates the number of progenitors (some of which might be outside this volume). For ROCKSTAR we show a second value in which only those with mass larger than $20m_p$ are considered.

absorbed.

If all the available halos are considered, ROCKSTAR is the catalogue with most small halos, leading to a higher branching ratio, which drops when removing the low mass halos. HBTHALO is also able to discern more substructure, yielding a slightly higher $N_{\rm dprog}$ than SUBFIND and AHF.

From the tree building point of view we remark that SUBLINK, with the possibility of omitting one snapshot, increases $N_{\rm dprog}$ considerably for the two catalogues with more substructure: ROCKSTAR and HBTHALO. HBTTREE, in modifying the catalogue, tends to recover the halo set generated by HBTHALO. This effect is more noticeable in the case of SUBFIND because it is also based on FoF catalogues (Section 1.2). JMERGE shows very little branching ($N_{\rm dprog} = 1$ or 2) because by construction it never associates a small merging halo with a much bigger one. It rather associates the in-falling halo with another small halo.

Note, however, that this was a very extreme case and that Figure 1.9 is not necessarily representative of the statistics seen in Figure 1.8, rather it helps to understand the kind of factors that influence the branching ratio.

1.5 Mass Evolution

The mass evolution of halos is an important input for semi-analytical models of galaxy formation. In this section we will study it through mass growth (Section 1.5.1) and fluctuations in mass (Section 1.5.2).

1.5.1 Mass Growth

Mass growth can be characterised by the discretised logarithmic growth, defined as:

$$\frac{\mathrm{d}\log\ \mathrm{M}}{\mathrm{d}\log\ \mathrm{t}} \approx \alpha_M(k, k+1) = \frac{(t_k + t_{k+1})(M_{k+1} - M_k)}{(t_{k+1} - t_k)(M_{k+1} + M_k)} \tag{1.10}$$

where k and $k+1$ are a halo and its descendant, with masses M_k and M_{k+1} at times t_k and t_{k+1}, respectively [64]. In order to reduce the range of possible values of this

variable to the finite interval $(-1, +1)$, we define:

$$\beta_M = \frac{1}{\pi/2}\arctan(\alpha_{\rm M}) \tag{1.11}$$

Figure 1.10 shows the distribution of β_M for three populations: all halos (A, on the left), main halos (B, in the centre) and subhalos (C, on the right). All distributions have been normalised by the total number of events found in halo sample A in each case. Selection is done as follows: all the halos identified at $z = 0$ are traced back along the main branch and at any snapshot if both a halo and its descendant are main [sub] halos and have mass $M > M_{th}^{\rm main}$ [$M > M_{th}^{\rm sub}$] (Table 1.1) sum to the population B [C]. The population A is compiled similarly, but taking all pairs of halos satisfying $M > M_{th}^{\rm main}$, regardless of being main or subhalos. Note that the distribution A is dominated by main halos, since they are more numerous.

Within the hierarchical structure formation scenario one expects halos to grow over time. This can be appreciated in column A, where the distribution of β_M is skewed towards values $\beta_M > 0$. However, there is a non-negligible number of cases ($\sim 15 - 30\%$) where it decreases ($\beta_M < 0$). While mass loss could be associated with tidal stripping of subhalos, column B shows that this is not the sole explanation within this simulation: while subhalos have an important contribution at the very far end of the distribution (corresponding to large mass losses), there are also many instances leading to $\beta_M < 0$ for main halos. Nevertheless, there are physical ways for main halos to lose mass: when two main halos approach each other, the effective radius for tidal stripping extends beyond the virial radius of the larger halo [see 108, for an elaborate discussion of exactly this phenomenon], thus, the small one can experience mass loss before becoming a satellite. Also, when halos change their shape, the specific halo mass definition (e.g. M_{200c} for AHF/Rockstar) of a halo finder can lead to an apparent mass loss.

The plot clearly shows that the differences across halo finders are greater than the variations introduced by the tree building method, with the exception of HBTtree (that modifies the input halo catalogue). There are two distinct classes of distribution for main halos (B): on the one hand, Rockstar and AHF, and on the other hand, SUBFIND and HBThalo which have a more skewed distribution. Recall from

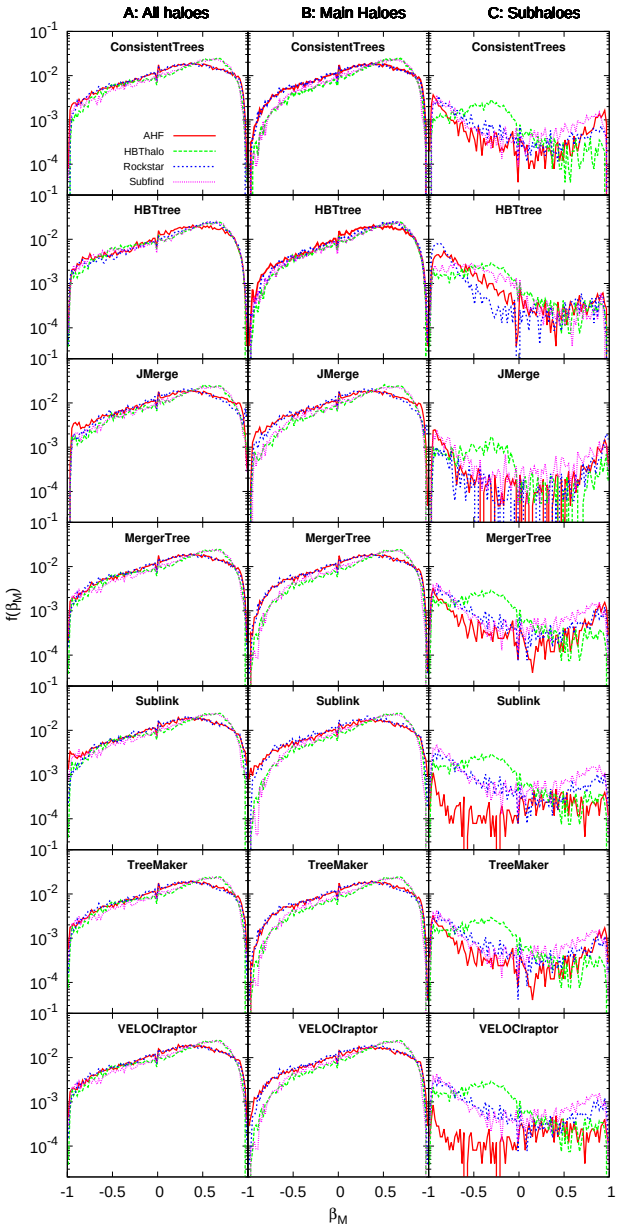

Figure 1.10: Mass growth distribution between two snapshots, β_M, related to the logarithmic mass growth through Equation 1.11, for halos that can be identified at $z = 0$, with mass $M > M_{th}$ at both output times. We distinguish 3 populations: A which contains all halos with $M_{th} = M_{th}^{main}$, B with only main halos and $M_{th} = M_{th}^{main}$, and C with only subhalos and $M_{th} = M_{th}^{sub}$. M_{th} is tabulated in Table 1.1 for the different halo finders. Each row displays a different tree building algorithm (as indicated). Each halo finder has its own line style as indicated in the legend. The distribution is computed as a histogram, normalised by the total number of events found by the corresponding halo finder for the population A.

(Section 1.2) that the former use an inclusive mass definition, thus, for a subhalo that just crossed the centre and is moving away, the total (inclusive) mass of the host halo can decrease if part of that subhalo crosses R_{200c}.

We finally remark that while subhalos are present in our somewhat low-resolution simulation (when compared to the state-of-the-art), they contribute significantly to neither the shape nor the amplitude of the mass growth distribution shown in column A (all halos). However, their own distribution (column C) is interesting in its own regard: we primarily observe mass loss due to tidal stripping, i.e. an imbalance of the distribution towards negative β_M values. In this case we find that whereas HBTHALO follows one distribution, the other three follow their own. This reflects the inconsistency in subhalo mass functions already seen in Figure 1.3.

In conclusion, most of the differences in the mass growth β_M can be accounted for by the choices made by the respective halo finder when defining quantities. In particular, HBTHALO and SUBFIND agree best with the *a priori* expectation from hierarchical structure formation.

1.5.2 Mass Fluctuations

After studying mass growth above, we quantify mass fluctuations by using

$$\xi_M = \frac{\beta_M(k, k+1) - \beta_M(k-1, k)}{2} \tag{1.12}$$

where $k-1, k, k+1$ represent consecutive time-steps. When far from zero, it implies a growth followed by a dip in mass ($\xi_M < 0$) or vice versa ($\xi_M > 0$). Within the hierarchical structure formation scenario this behaviour can be considered unphysical and equates to a snapshot where the halo finder might not have assigned the correct mass – though there are certainly situations where the definition of *correct mass* remains arguable. Nevertheless, it provides another means of quantifying the influence of the halo finder upon a merger tree.

The (normalised) distribution of ξ_M is presented in Figure 1.11 in the same way as Figure 1.10, i.e. three distinct columns for all halos (A, left), main halos (B,

middle), and subhalos (C, right). It reconfirms most of the claims of Section 1.5.1. We again find the distribution is essentially independent of the tree builder (besides HBTtree) for all three populations. We find two types of distributions for main halos (B): on the one hand, the SUBFIND and HBThalo catalogues give the broadest distributions and on the other hand, Rockstar and AHF have a more peaked distribution. This implies that the first pair of halo finders present more mass fluctuations ($\xi_M \neq 0$) than the second one. Note that this pairing is identical to the one reported in Section 1.5.1. And we also find (again) that subhalos (C) do not provide an explanation for the wings of the mass fluctuation distribution in column A, even though their own plot indicates that they predominantly undergo abrupt changes, i.e. they have easily distinguished wings.

Given that subhalos often undergo fluctuations (column C of Figure 1.11), this could cause fluctuations in main halos when the mass is defined exclusively (HBThalo and SUBFIND). In order to study this effect, we selected a halo whose mass evolution is characterised by a large ξ_M value (for the SUBFIND/HBThalo pair) in Figure 1.12. We localised the same object (the blue halo) and surrounding ones (red a green) in all four halo catalogues, showing the three consecutive snapshots used for the calculation of ξ_M given at the very right hand side of each panel. The halo undergoes a mass fluctuation for the finders HBThalo and SUBFIND, while it keeps growing for AHF and Rockstar. Figure 1.12 shows that, although it is true that for HBThalo/ SUBFIND the total mass of the subhalos increases when the main halo decreases and vice versa, the fluctuation of subhalo mass is one order of magnitude smaller than the main halo fluctuation and this cannot be the sole explanation. The fact that the red halo changes from being a subhalo to a main halo and then back to a subhalo again may be related (in a non-trivial way, since masses are defined exclusively) to the mass fluctuation. For this simple (compared to Figure 1.7 & Figure 1.9) configuration of halos, all the tree building algorithms agree in the resulting trees. We also note that even small fluctuations (10% in mass) are detected by this parameter ξ_M, in part due to an enhancement of ξ_M at late times (cf. Equation 1.11 & Equation 1.12).

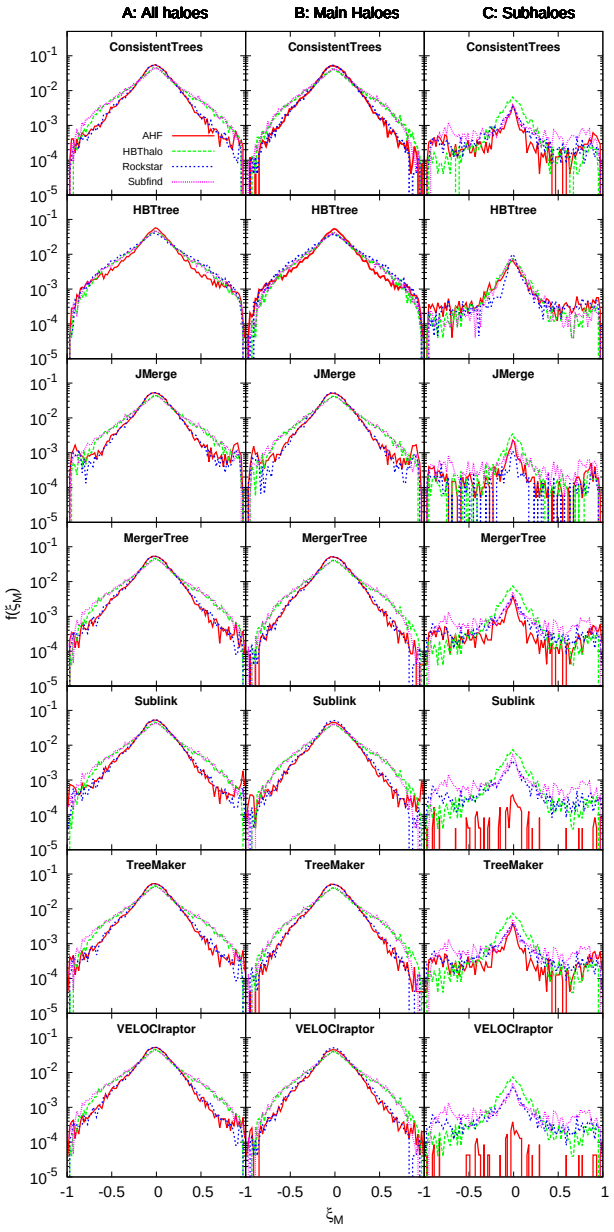

Figure 1.11: Distribution of mass fluctuations ξ_M (Equation 1.12), for halos found in three consecutive snapshots along a main branch that can be identified at $z = 0$, with mass $M > M_{th}$ for each appearance of the halo. We distinguish 3 populations: A which contains all halos with $M_{th} = M_{th}^{main}$, B with only main halos and $M_{th} = M_{th}^{main}$, and C with only subhalos and $M_{th} = M_{th}^{sub}$. M_{th} is tabulated in Table 1.1. Comparison is made between different tree builders (each row as labelled) and halo finders (line styles as in the legend). The distribution is computed as a histogram normalised by the total number of events for the corresponding halo finder for the population A.

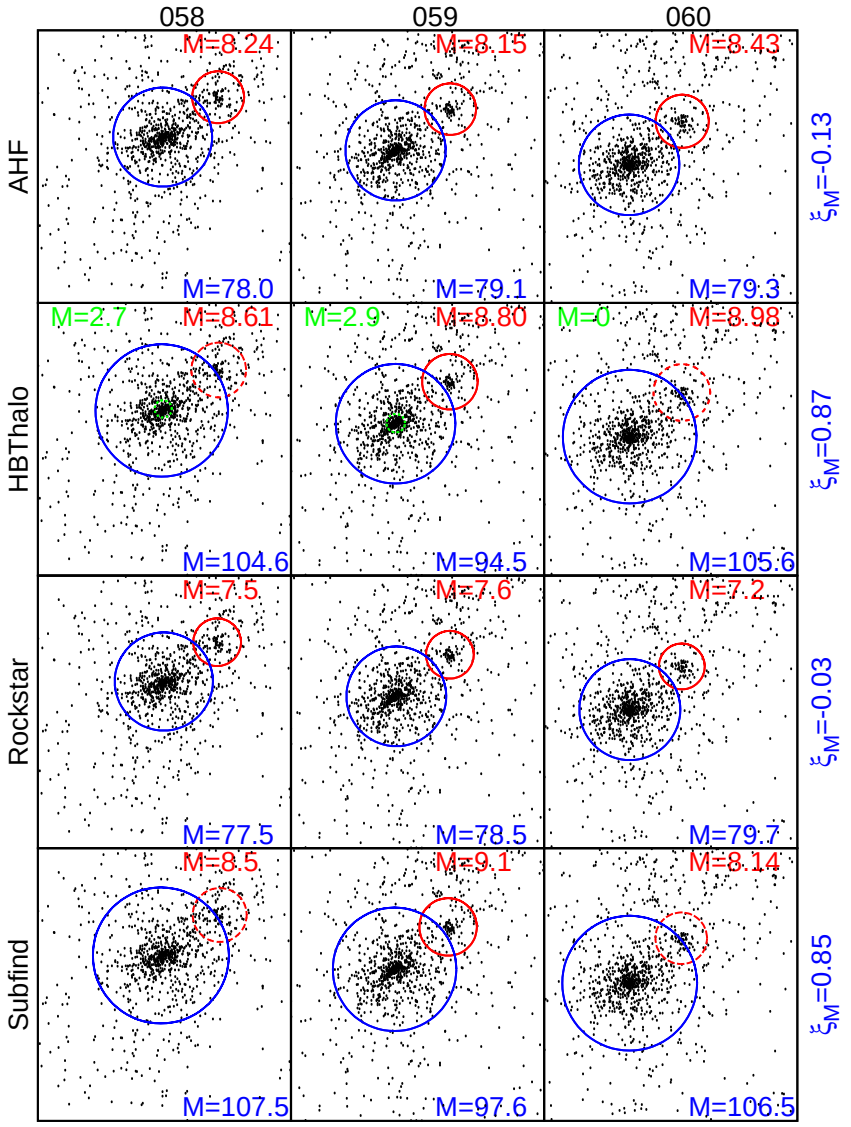

Figure 1.12: Projected 1 Mpc/h-side cube containing two halos (three for HBTHALO) evolving from snapshot 058 (left column) to 059 (central column) to 060 (right column). Each row shows a different halo finder. The radius of the circle is represented proportional to the mass of the object, with an extra factor of $\times 5$ for the small (red and green) halos. Dashed lines denote subhalos whereas solid lines are used for main halos. The mass of each halo is also shown in units of $10^{10}h^{-1}M_{\odot}$. At the right of each row we can see the value of ξ_M for the big halo, which quantifies the mass fluctuation as defined by Equation 1.12.

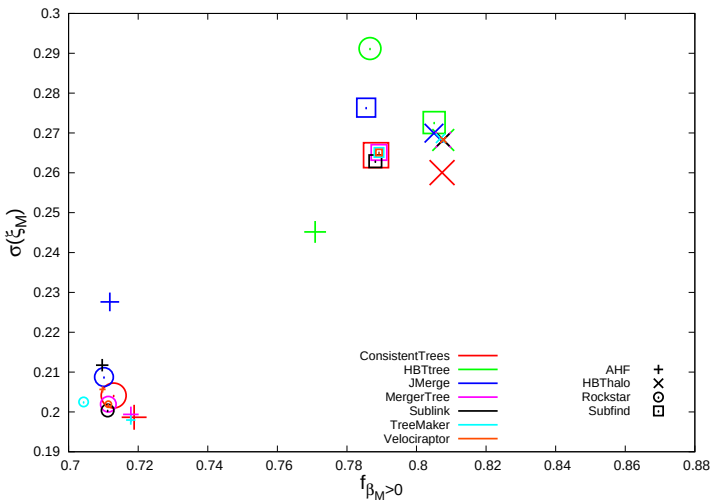

Figure 1.13: Summary of Figure 1.10 and Figure 1.11. On the abscissa we show the fraction of halos for which mass grows; on the ordinate we show the standard deviation of the mass fluctuations. Only main halos satisfying $M > M_{th}^{main}$ (Table 1.1) are taken into account. Every point represents a combination of a tree builder (size and colour-coded) and a halo catalogue (symbol-coded, see legend).

1.5.3 Combining growth and fluctuations

To better draw any conclusion from our study of the mass evolution of *main halos* we summarise results from their β_M and ξ_M statistics (Section 1.5.1 & Section 1.5.2) in Figure 1.13: the x-axis shows the fraction $f_{\beta_M>0}$ of objects for which $\beta_M > 0$, whereas the y-axis shows the standard deviation σ_{ξ_M} of ξ_M. Different sizes (or colours) now represent different tree building methods whereas the symbols stand for the input halo catalogue. The desirable feature of a tree describing hierarchical structure formation would be to have small mass loss for main halos (high $f_{\beta_M>0}$) and small mass fluctuations (low σ_{ξ_M}), at least *a priori*, because we also explained physical causes for these phenomena. Note also that the quantities plotted here do not provide a substitute for the whole curve shown in Figure 1.10 & Figure 1.11, but rather capture well the features of interest as they are observed. This summary plot illustrates very well how mass evolution sensitively depends on the choice of the halo finder:

- Points for the same halo finder (symbol) group together. The small scatter

amongst those groups represents the small influence of the tree building method on these magnitudes.

- HBTtree points deviate from the group, approaching the area of the HBThalo finder (crosses).

- The pair of halo finders HBThalo/SUBFIND achieves a lower rate of mass loss at the price of having more mass fluctuation than the other pair of finders AHF/Rockstar for main halos. We relate this pairing to the mass definition of the halo finder: the former is exclusive and uses self-bound objects, whereas the latter uses inclusive spherical M_{200c} objects.

We have verified that mass growth and fluctuations are intrinsically related to the mass definition. A simple change from an inclusive to an exclusive halo catalogue or from M_{200c} to arbitrarily shaped halos would change the shape of the curves seen in Figure 1.10 & Figure 1.11 and the position of the points in Figure 1.13. But other fundamental properties of the halo finder also leave their imprint, the evident differences between HBThalo and SUBFIND in Figure 1.13 are a proof of this.

1.6 Conclusions

We investigated the influence of the input halo catalogue on the quality of the resulting merger trees. 'Quality' in this regard has been identified as length of the main branch, number of direct progenitors, and, quantities that are highly relevant for semi-analytical modelling, the mass growth and mass fluctuation of halos. We also showed some specific examples of cases that aided our understanding of the influence of the halo finder and tree builder on the resulting properties of the trees.

In total, seven different tree building methods have been applied to the halo catalogues produced by four different halo finding algorithms which examined the same cosmological simulation. This produced 28 merger trees to be analysed. The influence of both groups of codes is summarised below, and the particular achievements and difficulties of the different methods discussed.

The influence of the halo finder

The primary conclusion of all the studies presented here is that the influence of the input halo catalogue is greater than the influence of the tree building method employed. This is especially clear for the mass evolution studies (Section 1.5) although it is also noticeable from the results of the main branch length (Section 1.4.1) and the studies on the branching ratio also suggest it (Section 1.4.2). Part of these differences are due to the fact that for this comparison we allowed the halo finders to choose their own definitions instead of unifying them as done in previous halo finder comparison projects. However, this way we find the real impact a user will encounter when choosing one or the other halo finder for his/her analysis.

Another pattern encountered along our studies is the pairing AHF/ROCKSTAR vs. HBTHALO/SUBFIND. This is very clear in the mass evolution of main halos (central columns of Figure 1.10 and Figure 1.11, summarised in Figure 1.13) and can also be seen in the main branch length distribution (Figure 1.5). We interpret this pairing to be caused by the fundamental construction of the halo catalogues, namely spherically truncated M_{200c} inclusive masses (Equation 1.9) for the former pair vs. self-bound exclusive objects starting from FoF groups for the latter. These differences can already be acknowledged in the main halo mass function shown in Figure 1.3.

The studies on the length of the tree (Section 1.4.1) are the cleanest test, since they do not rely on arbitrary choices such as the lower mass cut (which makes a significant difference for the branching ratio) or the mass definition (which is of great influence in the mass evolution). The tracking nature of HBTHALO showed excellent results in this section, with no early truncation of (sub)halos. ROCKSTAR and AHF showed early truncation of trees, especially for subhalos near the centre of their host, whereas SUBFIND did not show too much early truncation of subhalos, because they are systematically missing in the centre of the hosts. AHFled to the shortest main branches: halos disappear due to the high-z low-M incompleteness and the main branches tend to end early.

The relevance of the lower mass cut was also seen in the study of the branching ratio (Figure 1.8 in Section 1.4.2). In particular, for ROCKSTAR a cut in mass was not equivalent to a cut in the number of particles. Because of this, doing the same cut

in particles as for other catalogues, the branching ratio of ROCKSTAR was too high.

The mass evolution of halos was found to be mostly dependent upon the mass definition employed by the halo finder. However, it is not clear which finders perform best: HBTHALO/SUBFIND show less mass loss whereas AHF/ROCKSTAR show fewer mass fluctuations. Mass evolution is intrinsically related to the way the mass is defined, and the choice of a different mass definition within the same halo finder would lead to different results.

Along these lines, note that some properties of the halo finders are simple choices that are relatively easy to change, as for example the exclusive/inclusive mass assignment or the choice of spherical halos vs. self-bound objects. However, we have seen in [96] that other, more fundamental, details of each halo finder (such as the initial particle collection) leave their own unique signature in the catalogue. These are practically unavoidable and hence users have to decide upfront which halo finder best suits their needs.

The influence of the tree building method

Although we found a greater dependence on the halo finder than on the tree building method, each of the tree codes also has its own peculiarities:

- CONSISTENT TREES in many cases is able to correct the problems posed by the finder by adding artificial halos.
- HBTTREE, when recomputing the substructure, makes halos more traceable, improving the results.
- JMERGE has problems in dealing with the motion of (sub)halos in highly clustered environments.
- MERGERTREE, TREEMAKER and VELOCIRAPTOR behave very similarly, as they are based on nearly identical algorithms.
- SUBLINK is sometimes able to compensate for non-detection of halos by looking at non-consecutive timesteps.

In these lines, and confirming results from [64], being able to skip snapshots or having a tracking nature is found to be crucial in properly trace the history of halos.

Outlook

The main outcome of the present study is that the fundamental properties of halo finders have a major impact on the merger trees constructed from them, and that some tree building techniques can help improving those trees by correcting for halo finder defects. We pointed out the repercussions that several properties of the halo finders and tree building codes can have on the final trees. This should help the community choosing, designing or modifying their pipelines to construct merger trees idealised for their specific purposes.

It is worth mentioning that, although here we focused on the differences among the resulting merger trees, the agreement among them is nevertheless remarkable. The general features of the trees resulted as one would have expected, and are similar from one tree to another. Many times the differences between trees are only seen when plots are done on a logarithmic scale, since those differences are at the order of a few-cases for every thousand plotted.

The series of workshops and studies within the Mocking Astrophysics programme is helping us to quantify the degree of understanding that we actually have about structure formation. It helps the community validating and improving their algorithms used in the simulation pipeline, whose outcome we compare with observations to learn about the physics of the Universe. This process will continue, since for every milestone reached in Cosmology a few more arise on the horizon.

Chapter 2

HALOGEN: an approximate halo catalogue generator

2.1 Introduction

2.1.1 Approximate halo mock catalogues

We have entered an observational era where it is customary for redshift surveys to map millions of galaxies in the sky with the volumes of these surveys exceeding Gpc^3 scales. Recent and upcoming galaxy survey projects include PAU[1] [109], BOSS[2] [110], DES[3] [19], DESi[4] [111], Euclid[5] [20], LSST[6] [112] etc. The interpretation of such surveys demands a new generation of theory tools in order to better understand and interpret the large amounts of data. One important component is the need for accurate simulations of the expected results, to which the observations should be compared. However, models of large-scale-structure and the clustering of (dark matter) halos forming in it are inherently non-linear, and require the production

[1] http://www.pausurvey.org/
[2] http://Cosmology.lbl.gov/BOSS/
[3] http://www.darkenergysurvey.org/
[4] http://desi.lbl.gov/
[5] http://www.euclid-ec.org/
[6] http://www.lsst.org/

of simulations based on N-body calculations (as explained in Section 1.1.1). Such simulations are extremely costly, and consequently very few realisations can be run for a given application. However, investigating the effects of systematic errors, cosmic variance, and their interplay require many hundreds or even thousands of realisations of a single simulation (e.g. BOSS survey used 600 [113]). These are necessary to compute covariance matrices which characterise the resultant uncertainty on the final parameters.

To mitigate this situation, it is now a common practice to use approximate schemes in order to calculate the required realisations of the simulations. Early such work used the so-called log-normal realisations [114], which placed particles randomly according to a log-normal distribution, given the true power spectrum. While this is indeed efficient, and reproduces 2-point statistics faithfully, its lack of physical motivation for the particle placement results in poor higher-order statistics, such as the 3-point function or Counts-in-Cells moments. Improved methods developed in the past decade include PTHALOS [113, 115], PINOCCHIO [116, 117], PATCHY [118], COLA [119], QPM [120], EZMOCKS [121], FASTPM [122] etc.

One may segregate these methods into two classes – predictive-type methods which are required to 'find' halos in a given density field (e.g.. PINOCCHIO, COLA and PTHALOS), and statistical-type methods which merely stochastically sample a density field to locate halos (e.g.. PATCHY, QPM and EZMOCKS). The former have the advantage of being predictive, and often not requiring an N-body reference simulation for calibration, while the latter have the advantage of computational speed and resources, as the large-scale statistics are produced by an approximate gravity solver – including higher-order statistics. See Section 2.6 for a comparison based on [3].

In this chapter we present a (statistical-type) approximate scheme, called HALOGEN [2], whose prime objective is to generate halo catalogues with the correct 2-point clustering and mass-dependent bias using a simple and rapid approach.

We note that statistical-type methods tend to follow a standard pattern of four steps:

1. Produce a density field.

2. Sample halo masses.

3. Sample particles as halos with some bias.

4. Assign halo velocities.

In this study we seek to abstract this pattern, providing a framework in which each step is highly modular. Whilst modular, HALOGEN implements default behaviour with very simple (and rapid) components – using 2$^{\text{nd}}$-order Lagrangian Perturbation Theory (2LPT) as the gravity solver, theoretical mass functions, a single-parameter bias prescription (as opposed to 2 or more parameters for other statistical-type methods) and a direct linear transformation of the velocities. As such, HALOGEN can be rapidly calibrated, and easily extended. In addition, we introduce physically motivated constraints for *halo exclusion* and *mass conservation*, which tie the individual steps together.

We will compare the results from HALOGEN to the reference N-body simulations presented in Section 2.1.2. We introduce the general ideas of the method in Section 2.2, leaving a more detailed explanation of the spatial placement of halos – which we consider the essence of HALOGEN – for Section 2.3. Section 2.4 demonstrates the effects of each parameter of HALOGEN and how to optimise them. We present some applications and results of HALOGEN in Section 2.5 and compare it to other methods section Section 2.6

2.1.2 The reference simulations

To tune HALOGEN to a specific cosmology, we require an N-body simulation. In order to show the adaptability of HALOGEN to varying setups, we have not limited ourselves to a single simulation but used several references with differing box size, mass resolution, and cosmology. Further, the reference halo catalogues have been obtained by applying different halo finding techniques, and have different number density. We summarise the characteristics of both reference catalogues in Table 2.1 and describe them below.

Goliat Simulation This simulation was run with the GADGET2 code [60] from initial conditions generated by 2LPTIC[7] at $z = 32$. It uses $N = 512^3$ dark matter particles in a box with side length $L_{\rm box} = 1000h^{-1}$Mpc. The cosmological parameters used in this simulation are $\Omega_M = 0.27$, $\Omega_\Lambda = 0.73$, $\Omega_b = 0.044$, $h = 0.7$, $\sigma_8 = 0.8$, $n_s = 0.96$ yielding a mass resolution of $m_p = 5.58 \times 10^{11} h^{-1} M_\odot$. In this catalogue we use a reference halo number density of $n = 2.0 \cdot 10^{-4} ({\rm Mpc}/h)^{-3}$ The halo catalogue was obtained from a $z = 0$ snapshot and has been generated with the halo finder AHF [123], a spherical-overdensity (SO) algorithm (see Section 1.2). Though AHF identifies subhalos, they have been discarded for the present analysis as these scales are too small for 2LPT to resolve. We show in Section 3.2.3 how to add substructure in a phenomenological way following a Halo Occupation Distribution.

HALOGEN requires an input density field obtained from 2LPT (see Section 1.1.1). For this purpose, we run a 2LPTIC snapshot at $z = 0$ with the same initial condition phases as those used in GOLIAT.

Big MultiDark Simulation[8] BIGMULTIDARK described in [124], employs the cosmology from the Planck CMB mission [125], which for some parameters represents a significant change with respect to the GOLIAT simulation: $\Omega_M = 0.31$, $\Omega_\Lambda = 0.69$, $\Omega_b = 0.048$, $h = 0.68$, $\sigma_8 = 0.82$, $n_s = 0.96$. The halo catalogue is extracted with a Friends-of-Friends (FOF) [97] algorithm (which intrinsically neglects substructure) at $z = 0.56$, and we choose a reference halo number density of $n = 3.5 \cdot 10^{-4} (Mpc/h)^{-3}$.

Compared to GOLIAT, is both larger ($L_{\rm box} = 2500h^{-1}$Mpc) and more resolved ($N = 3840^3$ particles of mass $m_p = 2.3 \times 10^{10} h^{-1} M_\odot$). It was run with L-GADGET2 from initial conditions based on the Zel'dovich Approximation (ZA) at $z = 100$. Given the large scales that it explores while resolving large numbers of halos, it is well suited to probing the Baryon Acoustic Oscillation (BAO) peak.

For the input of HALOGEN we run 2LPTIC to $z = 0.56$ with the same initial condition phases as BIGMULTIDARK. The cosmology and L_{box} used are the same, but with a lower resolution of $N = 1280^3$.

[7] http://cosmo.nyu.edu/roman/2LPT
[8] http://www.cosmosim.org

Name	$L_{\rm box}$	$N_{\rm part}$	z	Ω_b	Ω_M	Ω_Λ	h	σ_8	n_S	Finder	n	IC	$z_{\rm IC}$
GOLIAT	1000	512^3, 512^3	0	0.044	0.27	0.69	0.7	0.8	0.96	AHF	$2.0 \cdot 10^{-4}$	2LPT	32
BIGMULTIDARK	2500	3840^3, 1280^3	0.56	0.048	0.31	0.73	0.68	0.82	0.96	FOF	$3.5 \cdot 10^{-4}$	ZA	100
MICE	3072	4096^3, 1280^3	0,0.5,1,1.5	0.044	0.25	0.75	0.7	0.8	0.95	FOF	$15.6 \cdot 10^{-4}$	ZA	100

Table 2.1: Properties of the three reference N-body halo catalogues. From left to right: Side-length of the simulated cubic volume (in h^{-1}Mpc), number of particles (for N-body and HALOGEN), redshift of the snapshot, cosmological parameters (density of baryons, total matter and dark energy, Hubble parameter, power spectrum normalisation and spectral index), halo finding technique, halo number density (in $(\mathrm{Mpc}/h)^{-3}$), method used to generate the initial conditions and redshift at which they were generated.

This simulation will also be the reference for the comparison of the methods in Section 2.6

MICE Grand Challenge Simulation[9] The MICE Grand Challenge simulation (from now MICE), described in [126–128], is based yet in a different cosmology: $\Omega_M = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_b = 0.044$, $h = 0.7$, $\sigma_8 = 0.8$, $n_s = 0.95$. Halos are found by FOF, the simulation have a larger box size ($L_{\rm box} = 2500h^{-1}$Mpc) and more particles (4096^3), yielding a slightly lower mass resolution of $2.9 \times 10^{10} h^{-1} M_\odot$.

This simulation will not be used during this chapter, in which we are developing HALOGEN, but will be a reference to HALOGEN in Chapter 3, where we apply the method in the context of the Dark Energy Survey. This simulation is more appropriate for DES studies, since it can be used to build a lightcone of an octant of the sky up to $z = 1.4$. For this, we need to fit the model at several redshifts ($z = 0, 0.5, 1.0, 1.5$) and we use a larger number density reference ($n = 15.6 \cdot 10^{-4}$ at $z = 0.5$). Again, the same initial condition phases were used for the calibration.

2.2 HALOGEN: the method outline

In this section we briefly outline our method, leaving a more detailed presentation of the actual modus operandi of HALOGEN for Section 2.3. The general algorithm consists of four (major) steps:

- generate a dark matter density field,
- draw halo masses by sampling a halo mass function,

[9] http://maia.ice.cat/mice/

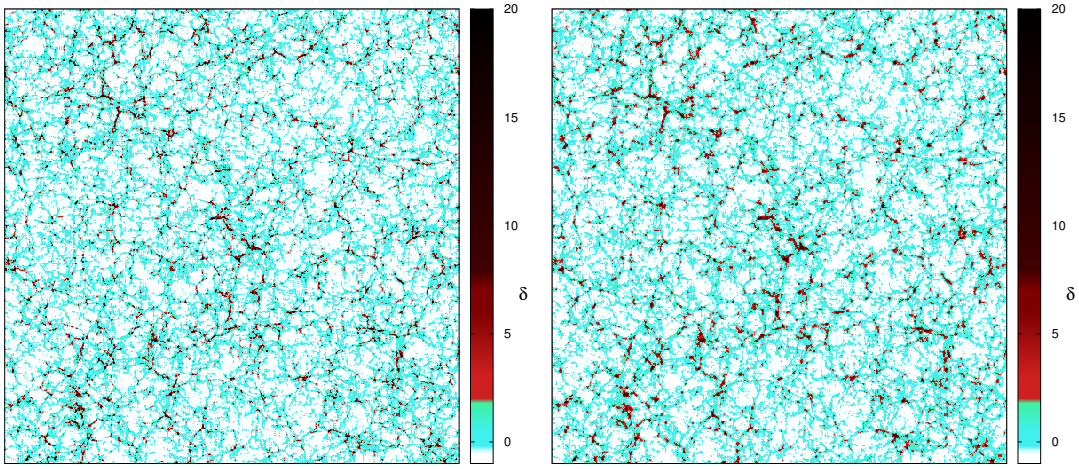

Figure 2.1: Here we show the difference between performing an actual N-body simulation (left) and using 2LPT (right) to generate a particle distribution at $z = 0.5$, with the same initial conditions. The image shows a slice of the density contrast δ distribution in a $1h^{-1}\text{Gpc}^3$ box.

- populate the volume with halos in the box, and
- assign velocities to the halos.

We aim to de-couple each of these steps from the others as far as possible so that different algorithms may be used at each point. The first two steps are relatively trivial, as they use pre-developed prescriptions from the literature, and we discuss these, and basic outlines of the last two steps, in this section.

2.2.1 Density Field

The basic scaffolding of HALOGEN is an appropriate dark matter density field realised at the desired redshift, sampled by N particles. For simplicity we choose to use 2^{nd}-order Perturbation Theory (2LPT) (see Section 1.1.1 or [49, 50]) to produce this field, which can be obtained with the public code 2LPTic.

We show in Figure 2.1 the density distribution of an N-body simulation (left panel)

and a 2LPT representation (right panel)at $z = 0.5$. Notably, the 2LPT distribution appears to be blurred in comparison to the N-body simulation. This is due to the fact that 2LPTic – as the name suggests – was originally designed only to generate initial conditions [129], since even 2^{nd}-order perturbation theory breaks down at low redshift when over-densities become highly non-linear. The small-scale difference in Figure 2.1 can be explained by shell crossing, an effect in which particles following their 2LPT trajectories cross paths and continue rather than gravitationally attracting each other in a fully non-linear manner [130, 131]. In order to compensate for shell-crossing, [113] advocates the use of a smoothing kernel over the input power spectrum. We tested the effect of this smoothing in HALOGEN but did not find any improvement in the final catalogue.

Nevertheless, 2LPT provides a suitable approximation of the large scale distribution of matter, where perturbations have not yet entered into the highly non-linear regime and this is sufficient for HALOGEN. Note that HALOGEN is in principle agnostic about the method in which this density field snapshot is produced. Other methods, for instance the "Quick-PM" COLA[119] or 3LPT could equally be employed by the user. A different choice of density field will yield somewhat different results, especially at smaller scales. As long as the chosen method reconstructs large scales correctly, the remaining steps of HALOGEN should be unmodified.

Despite this, we have by default incorporated 2LPTic as part of the HALOGEN code (which bypasses the costly I/O of writing the snapshot to disk), but also allow the user to provide an arbitrary snapshot with a distribution of N particles in a cosmological volume. Our choice for 2LPT was mainly driven by its low computational cost and success in the distribution of matter at large scales. We use this approach for all results in this study.

2.2.2 Halo Mass Function

The halo mass function (HMF) $n(> M)$ measures the number density of halos above a given mass scale. It is required to generate mass-conditional clustering, which in turn is a pre-requisite for extension to HOD-based galaxy mock generation.

The most accurate HMF for a given cosmology, over a range of suitable scales, may be obtained from an N-body simulation via a halo-finding algorithm –although there are notable variations depending on the technique [81]. Since we require a full N-body simulation for the tuning of HALOGEN, it would be perfectly acceptable to use this simulation to generate the HMF. However, in the hope of future improvements, we wish to avoid using the full simulation as far as possible. Fortunately, there is a wealth of literature concerning accurate predictions of the HMF for widely varying cosmologies and redshifts using Extended Press-Schechter theory [39, 132].

The mass function may be calculated by any means, so as long as a discretised function of $n(> M)$ is provided. For simplicity, we decided to use the online halo mass function calculator HMFCALC[10] [133] for obtaining the halo mass distribution in this study. We produce a sampled mass function by the standard inverse-CDF method, utilising an arbitrary input HMF and sampling it as follows:

1. Either directly set a number density n_0 or compute it from a chosen minimum halo mass $M_{\rm min}$, i.e. $n_0 = n(> M_{\rm min})$.

2. Estimate the number of halos in the simulation volume V, i.e. $N_{\rm h} = n_0 \cdot V$.

3. Generate $N_{\rm h}$ random numbers y_i in the interval $y_i \in [0, n_0]$.

4. Invert the cumulative mass function ($n(> M)$) to obtain the halo masses $M_i = n^{-1}(y_i)$.

In Figure 2.2 we demonstrate how well the input HMF is reproduced, only differing from the mass function fit [Watson, 134] at high mass due to Poisson shot-noise controlled by the volume V (where expected numbers are of order unity). Further, the HMF of BIGMULTIDARK shows similar behaviour, indicating that the chosen fit is appropriate for this simulation.

[10] http://hmf.icrar.org

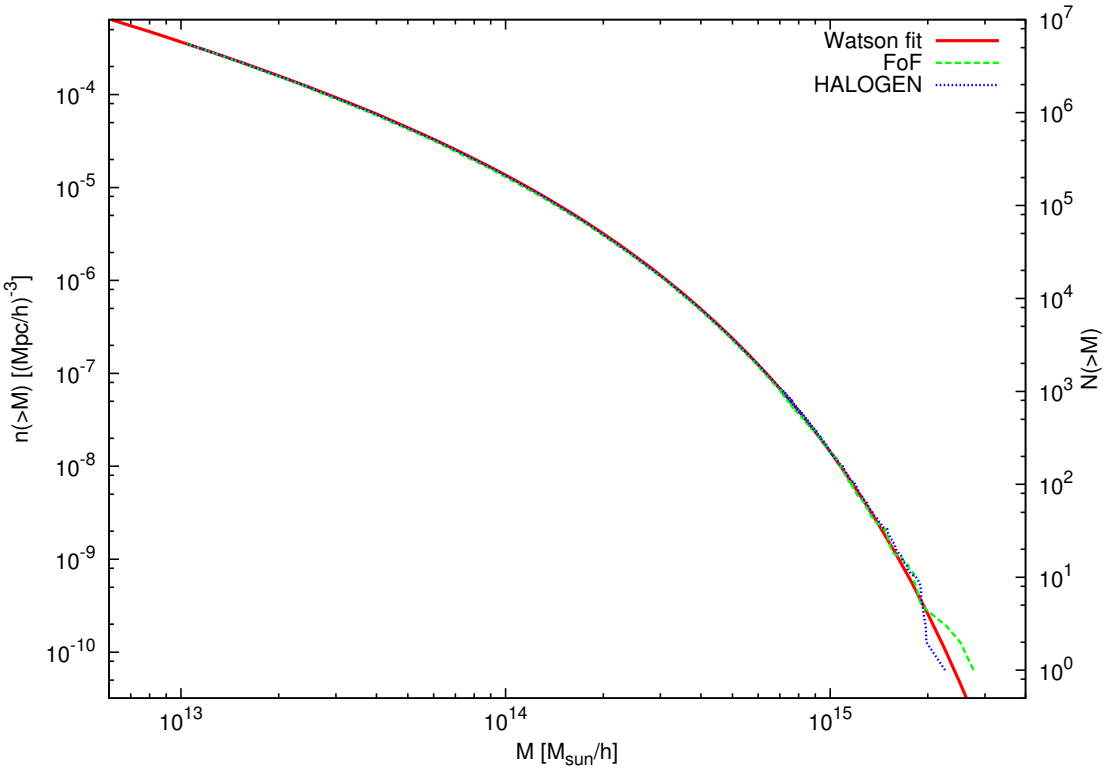

Figure 2.2: Mass function reconstruction by HALOGEN, as explained in the text. We show the theoretical fit $n_{\rm wat}(>M)$ as the solid red line, the resulting mass function of HALOGEN in a blue dotted line, and the mass function from the BIGMULTIDARK simulation in a green dashed line. The left y-axis represents the cumulative HMF in number density whereas the right y-axis has been multiplied by the volume $V = (2.5Gpc/h)^3$ and represents the actual number of halos above M. HALOGEN reproduces the HMF extremely well, as expected, though we note small fluctuations at the high-mass end due to discreteness effects.

2.2.3 Spatial placement of halos

The crucial step in the generation of approximate halo catalogues is the commissioning of halo positions. In keeping with the philosophy of modularity, the halo-placement step is de-coupled from the rest. Any routine which takes a vector of halo masses and an array of dark matter particle positions and returns a subset of those positions as the halo locations is acceptable. However, we consider this step to be at the heart of the HALOGEN method, as it is responsible of generating the correct mass-dependent clustering.

To achieve an efficient placement that reconstitutes the target two-point statistics, we recognise the validity of the clustering on large scales from the broad-brush 2LPT field. We place halos on 2LPT field particles, essentially using the estimated density field as scaffolding on which to build an approximate halo field. We will follow a series of steps in the construction of the method of spatial placement to be presented in Section 2.3 below.

2.2.4 Assignment of velocities

The most obvious way to assign velocities to each halo would be to use the velocity of the particle on which it is centred. However, halos are virialised systems whose velocities tend to be lower than that of their constituent particles. This is potentially mitigated by using the average velocity of all particles within a defined radius of the artificially placed halo. However, this is not robust as there are often very few particles inside the halo radius. Additionally, the 2LPT particle velocities will differ from their N-body counterparts due to shell-crossing, especially on the small scales associated with halos.

Thus, we prefer to take a phenomenological approach, and assume that a simple mapping via a factor $f_{\rm vel}$ can be applied to the collection of halo velocities to recover the results of the N-body distribution

$$\vec{v}_{\rm halo} = f_{\rm vel} \cdot \vec{v}_{\rm part}. \tag{2.1}$$

This factor could *a priori* depend on the velocity (i.e. a non-linear mapping) and the mass of the halo $f_{\rm vel}(v_{part}, M_{\rm halo})$. However, we will show in Section 2.4.2 that a linear mapping is sufficient and present a way to compute $f_{\rm vel}(M_{\rm halo})$.

2.3 HALOGEN: Bias scheme

Though HALOGEN is a four-stage process, the most crucial aspect is the assignment of halo positions, which this section describes in some detail. The general concept is to specify a sample of particles from an underlying density field as halos.

The motivating philosophy of HALOGEN is to start from the simplest idea and improve if necessary. In this vein, we present here successive stages of evolution of the HALOGEN method, which we hope will show satisfactorily that the method as it stands is optimal. Figure 2.3 will serve as the showcase for the various stages of HALOGEN. In it we present the 2-point correlation function (2PCF) for each stage of development to verify that the method approaches the GOLIAT reference catalogue as new characteristics are added.

Note that the 2PCF is computed with the publicly available parallel code CUTE[11] [135]. In the fitting routine that is included in the HALOGEN package and described in Section 2.4.1 we also use the same code.

2.3.1 Random particles

We start with the simplest approach: using *random* particles from the 2LPT snapshot as the sites for halos. We expect to recover the large-scale shape of the 2PCF in this way, as this is encoded in the 2LPT density field which we trace.

However, it is clear from Figure 2.3 that this method ('random no-exc') consistently underestimates the 2PCF over all scales except $r < 1h^{-1}$Mpc, where it should sharply drop to -1, but rather remains positive.

The consistent under-estimate is a realisation of an inaccurate linear bias, b, defined

[11] http://members.ift.uam-csic.es/dmonge/CUTE.html

as the scaling factor between the 2-point function of the halos and the underlying matter density field:

$$\xi_{\rm halo}(r) = b^2 \xi_{\rm dm}(r) \tag{2.2}$$

We begin to address this in Section 2.3.3.

The small-scale clustering can be explained by the fact that particles can be arbitrarily close, whereas distinct halos – recall that subhalos have been removed – have a well-defined minimum separation (otherwise they merge). The turn-over in the simulation based 2PCF occurs around the mean halo radius scale.

2.3.2 Random particles (with exclusion)

The simplest improvement to the random case is to eliminate the artificial small-scale correlations. Though the primary application of HALOGEN will be for large scales, a simple improvement at small scales is useful.

As we have noted, the artificial clustering at small scales arises from the fact that particles can be arbitrarily close, whereas simulated halos have a minimum separation. The radius of a halo is a rather subjective quantity, and its definition is modified in various applications and halo-finders. However, we may parametrise this by

$$R_\Delta = \left(\frac{3M_{\rm halo}}{4\pi \Delta_h \rho_{\rm crit}} \right)^{1/3}, \tag{2.3}$$

where Δ_h is the overdensity of the halo with respect to the critical density of the Universe. For the work presented here we used $\Delta_h = 200$.

Using this scale, we introduce *exclusion*, a modifiable option which controls the degree to which halos can overlap, which we set to mimic the halo finder's specification. For example, in this work we use both AHF and FOF (see Section 1.2). For the latter we do not allow any overlap whereas for the former HALOGEN's halo centres are not allowed to lie inside another halo's radius.

The effect of *exclusion* is presented in Figure 2.3 ('random exc'). As expected, scales of $r < 1h^{-1}$Mpc show a turnover while larger scales are unaffected. We note that

the turnover is at smaller scales for HALOGEN than for AHF. This is to be expected, as it is unlikely to find two AHF halos separated by a distance slightly exceeding R_Δ, due to reasons akin to the FOF over-linking problem. In such cases, there is an increased likelihood of the two halos being subsumed into one, or one becoming a subhalo of the other. It is conceivable that one could empirically model these effects by tuning the value of Δ_h by some factor which captures this suppressed probability. However, as we are more interested in large scales and these considerations touch upon the subtleties of halo definition, we consider these exclusion criteria sufficient for present purposes. We will use this form of exclusion (in an appropriate form) for all following work.

2.3.3 Ranked approach

We return now to the problem of under-estimation of the correlations, which we noted was due to an incorrect realisation of the linear halo bias. In effect, a random choice of particle position corresponds to sampling the matter power spectrum uniformly, and therefore $b = 1$. However, halo bias is generally greater than unity (especially for higher mass halo samples) [136].

Increasing the bias corresponds to sampling higher-density regions. The simplest way to achieve this is to rank-order the density of regions in the particle distribution, and assign halos to these regions based on their mass.

To calculate densities from the particle distribution, we simply create a uniform grid with cell-size $l_{\rm cell}$, and obtain the density in each cell using a Nearest-Grid-Point (NGP) assignment scheme [137]. We consider specification of the optimal $l_{\rm cell}$ in Section 2.4.3. The cells are ordered by density, and the halos by mass, and each halo is assigned to its corresponding cell (a random particle is chosen within the cell).

Using $l_{\rm cell} = 5h^{-1}$Mpc in this case, we obtain the results shown in Figure 2.3 labelled 'ranked exc'. The resulting 2PCF is now overestimated. This is not surprising, since even if we expect halos to form in dense environments, the bias is not completely deterministic: in reality the $n^{\rm th}$ most massive halo does not need to reside in the $n^{\rm th}$ densest place.

The effect of introducing a scale length, $l_{\rm cell}$, is also clearly seen in this result. There is a turnover in the 2PCF below $l_{\rm cell}$, which corresponds to a significant reduction of bias on these scales since a random particle is chosen within the cell.

2.3.4 α approach

We find that selecting completely random particles yields too low a bias, whereas the ranked approach is highly biased. We require an intermediate solution, which has higher probability of selecting dense areas than the random approach, and lower probability than the ranked approach.

The probability that a cell is chosen is a function of its density,

$$P_{\rm cell} \propto G(\rho_{\rm cell}). \tag{2.4}$$

In the completely random case, we have $G(\rho_{\rm cell}) = \rho_{\rm cell}$. In principle we can tailor $G(\rho_{\rm cell})$ so that the probability of selecting a cell reproduces the appropriate bias. We choose to constrain $G(\rho_{\rm cell})$ to have a power-law form, i.e.

$$G(\rho_{\rm cell}) = \rho_{\rm cell}^{\alpha}. \tag{2.5}$$

When $\alpha = 1$, we recover the random approach, and as $\alpha \to \infty$ we obtain the ranked approach.

In Figure 2.3 we show results for $\alpha = 1.5, 2$, demonstrating the effectiveness of our model for tuning the normalisation (i.e. bias) of the 2PCF. The $\alpha = 1.5$ curve closely matches the 2PCF of the AHF catalogue, at least at scales larger than the applied cell size $l_{\rm cell} = 5h^{-1}$Mpc.

The exact value of α for a particular application may be determined by a least-squares fit, which we describe in more detail in Section 2.4.1 (note that here the choice of α was not formally fit).

In corollary with this prescription, we also introduce a means to roughly ensure *mass conservation* in cells: once a halo is placed, if the total halo mass in the cell exceeds the original mass, the cell is eliminated from future selections. However, we

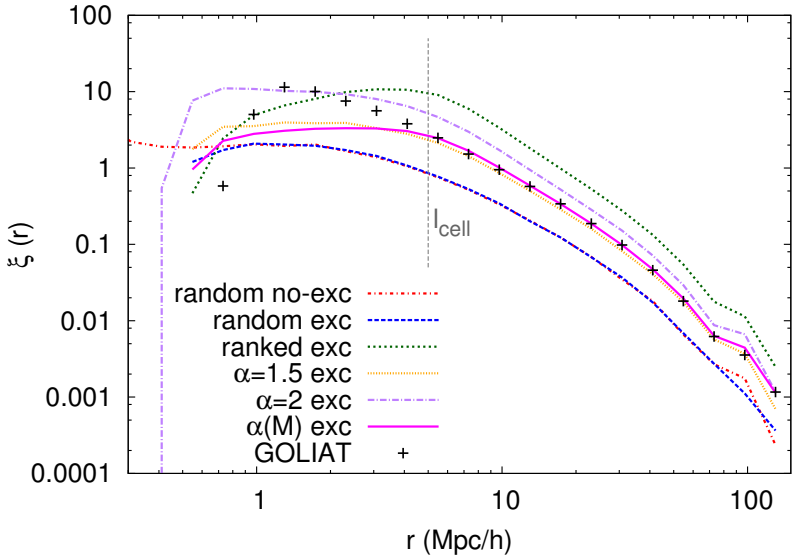

Figure 2.3: Two-point correlation function of the GOLIAT halos in comparison to HALOGEN for the various evolutionary stages presented in Sections 2.3.1 through 2.3.5. The dashed vertical line indicates the cell size of $l_{\rm cell} = 5h^{-1}$Mpc applied for the approaches 2.3.3 through 2.3.5.

do not update the value of the probability after every halo placement because it is computationally very expensive ($O(N_{\rm cell}^3)$) and we have checked that doing so has a negligible effect on output statistics.

We note that a similar method was employed in QPM [120]. In fact, the physically meaningful distribution is $f_{\rm halo}(\rho)$ – the fraction of halos in cells with density ρ. This can be written as

$$f_{\rm halo}(\rho) = P({\rm cell}|\rho) f_{\rm cell}(\rho), \tag{2.6}$$

where $P({\rm cell}|\rho)$ specifies the relative probability of choosing a cell given its density (in our case, ρ^α), and $f_{\rm cell}(\rho)$ is the intrinsic distribution of cell densities given the cell size and cosmology (heavily related to the cosmological parameter σ_8). QPM specifies the target distribution $f_{\rm halo}(\rho)$ directly, as a Gaussian. In HALOGEN we instead specify $P({\rm cell}|\rho)$, which is more closely tied to our algorithm. In principle one can convert from QPM-like methods to HALOGEN with Equation 2.6.

2.3.5 $\alpha(M)$ approach

The approach as it stands reproduces the 2PCF accurately down to the scale of $l_{\rm cell}$. If the 2PCF of a sample of given number density is all that is required for a specific application, then this will do well.

However, if we were to select a sub-sample of the most massive halos of our catalogues and recompute the 2PCF, the bias would be incorrect, since more massive halos are more biased [136]. For a truly representative catalogue, in which the halos are conditionally placed based on their mass, the bias model is required to be mass-dependent. Failing this, there is no physical meaning attached to the assignment of masses in the second step (Section 2.2.2).

Mass-dependent halo bias is also crucial for implementing HOD models on the catalogue, for use in galaxy survey statistics, as the number of galaxies associated with a halo depends on its mass.

We incorporate this mass-dependence into the α parameter, so that we finally have

$$G(\rho_{\rm cell}, M) = \rho_{\rm cell}^{\alpha(M)}, \tag{2.7}$$

with $\alpha(M)$ an increasing function.

In practice, we use discrete mass bins, and for each bin i, with masses $M_{\rm th}^{i-1} > M > M_{\rm th}^{i}$, we use a different α_i. We describe how we obtain the best-fit to this mass-dependent α using the fiducial halo catalogue from the simulation in Section 2.4.1.

Using just five mass bins, we illustrate this approach in Figure 2.3, labelled "$\alpha(M)$ exc" (magenta line) using the best-fit values for $\alpha(M)$. We list in Table 2.2 the mass thresholds, applied α-values, and corresponding number densities of all halos with $M_{\rm halo} > M_{\rm th}^{i}$. Note that though the probability is not recomputed after placing a halo, it is recomputed with updated ρ and α when changing mass bins.

Though the $\alpha(M)$ approach does not improve the 2PCF with respect to the α approach in Figure 2.3, it has the clear advantage of reproducing a mass dependent clustering, which as we noted is essential for further HOD analyses, and useful for being able to use any mass-range in the same realisation.

bin	$M^i_{\rm th}$ $[h^{-1}M_\odot]$	n_i $[(h^{-1}{\rm Mpc})^{-3}]$	α_i
0	$1.64 \cdot 10^{14}$	$0.05 \cdot 10^{-4}$	3.54
1	$4.80 \cdot 10^{13}$	$0.40 \cdot 10^{-4}$	2.26
2	$2.65 \cdot 10^{13}$	$0.90 \cdot 10^{-4}$	1.77
3	$1.86 \cdot 10^{13}$	$1.40 \cdot 10^{-4}$	1.48
4	$1.38 \cdot 10^{13}$	$2.00 \cdot 10^{-4}$	1.41

Table 2.2: Properties of the selected mass bins for the GOLIAT simulation: mass threshold $M^i_{\rm th}$, equivalent number density $n(M > M^i_{\rm th})$ and best fit α_i in $M^{i-1}_{th} < M < M^i_{th}$ for the HALOGEN $\alpha(M)$ approach.

2.3.6 Summary

In conclusion, HALOGEN constitutes a method for generating a halo catalogue which exhibits correct 2-point clustering statistics, while not only positioning the halos correctly, but also imbuing them with physically meaningful masses. The method can be summarised as follows.

The particles generated by 2LPT (Section 2.2.1) are covered by a grid of cell size $l_{\rm cell}$, the halo masses M_i generated from the halo mass function (Section 2.2.2) are ordered by mass, and starting from the most massive halo they are placed by

1. selecting a cell with probability $P_{\rm cell} \propto \rho^{\alpha(M)}_{\rm cell}$,

2. randomly selecting a particle within the cell and using its coordinates as the halo position,

3. ensuring that the halo does not overlap (following an *exclusion* criterion) with any previously placed halo in any cell, and re-choosing a different random particle in that case,[12]

4. subtracting the halo's mass from the selected cell, $m_{\rm cell} = m_{\rm cell} - M$: if $m_{\rm cell} \leq 0$ the cell is removed from selection.

[12] If, after several iterations all the particles are found inside another halo, re-choose cell (to avoid infinite loops).

Parameter	Motivation	Value
α_i	linear bias	χ^2-fit to bias
$f^i_{\rm vel}$	velocity bias	$f^i_{\rm vel} = \sigma^i_{\rm NB}/\sigma^i_{\rm p}$
$l_{\rm cell}$	algorithm	$l_{\rm cell} \approx 2 \cdot d_{\rm p}$

Table 2.3: A summary of the parameters involved in HALOGEN, the motivation to introduce them and how to compute/optimise them. See text for details

Note that the physically motivated nature of the process suggests that higher-order statistics may also be recovered with some success.

2.4 HALOGEN: Parameter Study

We have mentioned several parameters of the HALOGEN method, and these are of particular importance in producing accurate realisations. In this section we will discuss each parameter, its effects and how to optimise for it if possible.

There are three parameters in HALOGEN (with other options and parameters being expressly determined by the required output, such as the size of the simulation box L): the two physical parameters of the model, α – controlling the linear bias – and $f_{\rm vel}$ – controlling the velocity bias – and the one parameter of the algorithm, $l_{\rm cell}$.

In the previous Section we used GOLIAT as a reference. We now turn to BigMultiDark and its FOF catalogue: this simulation has a larger volume, allowing us to probe BAO scales. The increased volume also reduces cosmic variance on intermediate scales. HALOGEN primarily aims at reproducing clustering statistics for even larger volumes, hence it is beneficial to assess the performance of HALOGEN and its parameters in this regime. Furthermore, this demonstrates independence from the underlying simulation and halo finding technique.

2.4.1 Fitting $\alpha(M)$

The value of $\alpha(M)$ is crucial to the performance of HALOGEN, as it constitutes the only physical parameter controlling the bias. The HALOGEN package contains a

stand-alone routine which determines a best-fit for $\alpha(M)$, which can then be passed to HALOGEN to generate any number of realisations. We describe this routine here, and illustrate it with application to BIGMULTIDARK. The fitting of $\alpha(M)$ is based on the standard χ^2-minimisation technique. However, a few details are worth mentioning.

Mass-dependence. We perform the fit in sharp-edged mass bins to determine a mass-dependent $\alpha(M)$, i.e. for each bin i we fit a α_i for the mass range $M_{\rm th}^{i-1} < M < M_{\rm th}^{i}$. There are two conceivable ways of doing this – differentially or cumulatively. We have experimented with both and find that the cumulative procedure has better performance. That is, we fit the first mass bin, and then the first and second together (keeping the best value of α_0 for the first bin), and so on. This has the advantage of being able to properly correct for deviations in previous bins, which is particularly important since the first bins to be fit are the high masses, for which fewer halos exist. Misestimation of α here is more likely, but is compensated for when fitting to lower mass bins by including the high-mass estimates in the fit.

HALOGEN variance. The halo placement in HALOGEN is probabilistic, even given a constant underlying density field. Using different random seeds can slightly affect the final placement, and thus the clustering statistics (the extent of this is dependent on the volume, n and α). We term this "HALOGEN variance", and note that it is not to be confused with cosmic variance. Cosmic variance is introduced by modifying the the random seed of 2LPTIC, which in effect results in a different realisation of the universe[13]. During the fit each mass bin is realised several times (ten in the case of BIGMULTIDARK) with HALOGEN to average out the effects of HALOGEN variance, and also provide an error $\sigma_{\rm H}$ (computed as the standard deviation) to use in the definition of χ^2.

[13]Cosmic variance – strictly speaking – requires the study of the same volume, but in a different place in the universe. This approach is more appropriately called 'sampling variance' yet nevertheless the generally accepted technique for generating covariance matrices.

χ^2 minimisation. The fit is performed by minimising χ^2:

$$\chi^2(\alpha) = \sum_j \left(\frac{\xi_{\rm H}(r_j|\alpha) - \xi_{\rm NB}(r_j)}{\sigma_{\rm H}(r_j|\alpha)} \right)^2 \tag{2.8}$$

where $\xi_{\rm H}$ and $\xi_{\rm NB}$ are the 2PCFs of HALOGEN and the reference catalogue, respectively. We note that minimising this statistic is susceptible to systematic errors in HALOGEN in bins where the stochastic error (σ_H) is much smaller than the systematic error ($\Delta\xi$). This is especially likely when the region of the fit approaches $l_{\rm cell}$. To test whether the region is stable, we may choose a distance estimator to be minimised that treats all scales with the same weight, e.g.. $\Delta = (\xi_{\rm H} - \xi_{\rm NB})^2/\xi^2_{\rm NB}$. We have tried both definitions in our fitted range, and the results are left unchanged, indicating that the range of the fit is stable.

We use a grid of α to cover the expected result for each mass bin. We use a cubic spline interpolation over $\chi^2(\alpha)$ to locate a precise minimum for the best-fit α.

Number of mass bins. The number of bins to use in this procedure will depend on the needs of the user, and the size and resolution of the reference simulation. It determines the reliability of the mass-dependent clustering. For BIGMULTIDARK we distribute the halos into 8 roughly equi-numbered bins with the mass thresholds $M^i_{\rm th}$ as shown in Table 2.4. In that table we also show the best-fit α_i, and the equivalent number density n_i for each mass threshold.

Fitting Range. We restrict the range of the fit to scales in which the shape of $\xi_{\rm H}(r)/\xi_{\rm NB}(r)$ is flat. This corresponds to mid-range scales of $15h^{-1}$Mpc $< r <$ $47h^{-1}$Mpc, which avoids small-scale effects of HALOGEN, and large-scale cosmic variance.

The 2PCFs for our 8 values of n_i are shown in Figure 2.4, where we compare the results from HALOGEN against the BIGMULTIDARK reference catalogue. The range used during the fitting procedure and for the χ^2-minimisation is indicated by the vertical lines.

We note that the choice of α finely controls the bias. This is demonstrated in

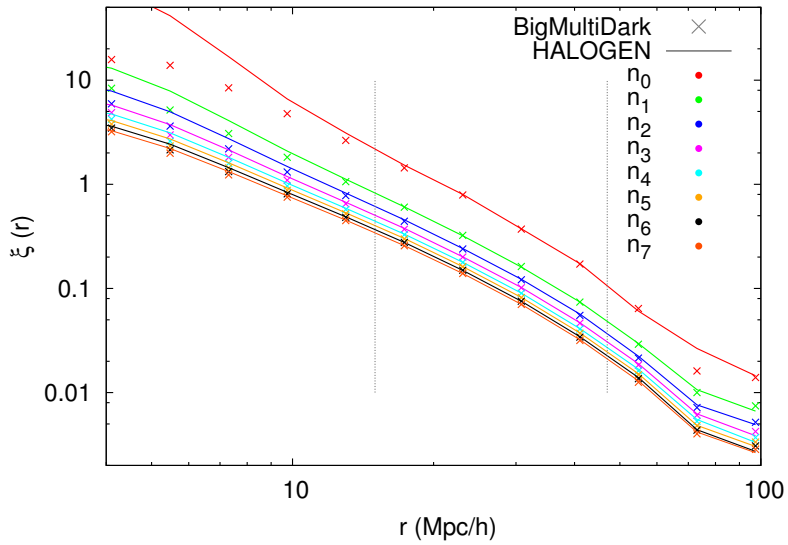

Figure 2.4: Correlation function of both BIGMULTIDARK (crosses) and HALOGEN halos (lines). We select 8 number densities n_i (colours in the legend) of halos, with values found in Table 2.4. The vertical dashed lines indicate the range of the fit.

Figure 2.5, in which we show the resultant $\xi(r)$ for the entire grid of α_7 for this fit (top figure). There is a $\sim$ 10 per cent deviation in $\xi_{\rm H}(r)$ over the grid range (1% between consecutive lines). On the bottom figure, we show the χ^2 of each of those curves and the cubic spline fit interpolation used to find the minimum, which corresponds to the α_7 best-fit value shown in Table 2.4.

2.4.2 Velocity factor $f_{\rm vel}$

In Section 2.2.4 we outlined a method of converting the velocity of 2LPTIC particles (designated as halo sites), $\mathbf{v}_{\rm p}$, to the velocity of a HALOGEN halo, $\mathbf{v}_{\rm h}$. We stated that the transformation was linear in $\mathbf{v}_{\rm p}$, and thus we can write

$$\mathbf{v}_{\rm h} = f_{\rm vel}(M) \cdot \mathbf{v}_{\rm p}, \tag{2.9}$$

where we have retained a mass-dependence in the conversion factor. This section will explore the means to calculate this factor.

We begin by justifying our choice of a linear function. Figure 2.6 shows the one-

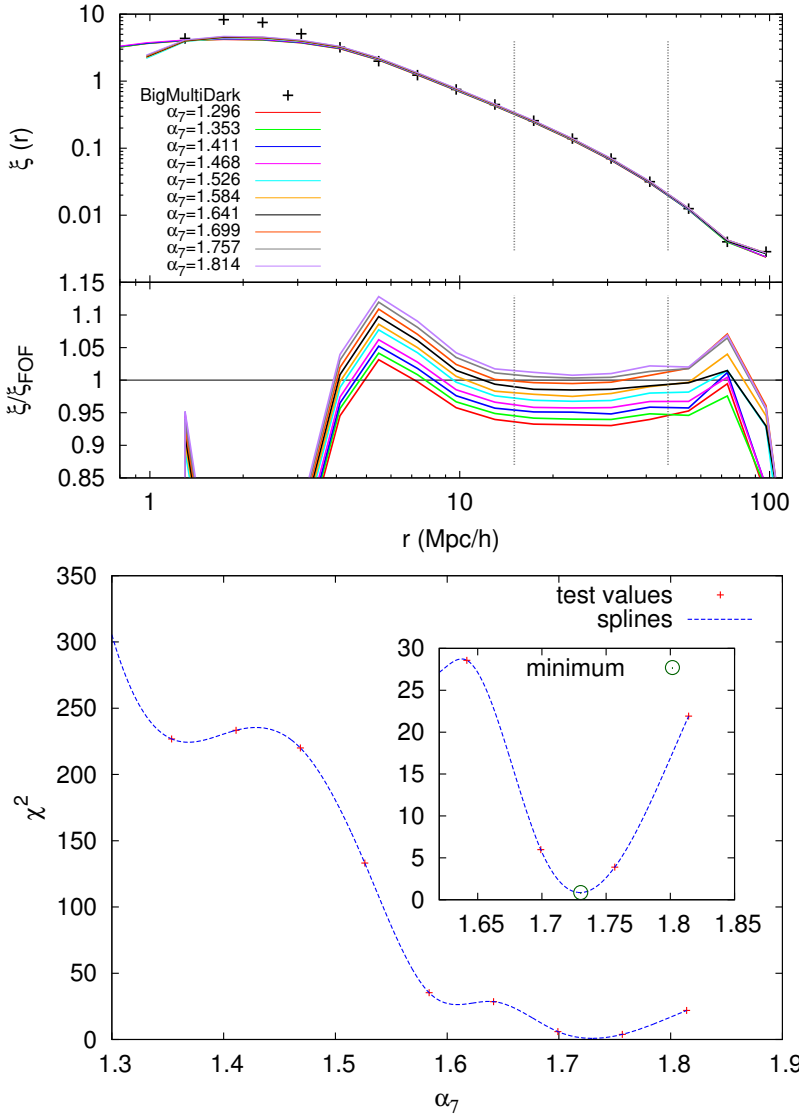

Figure 2.5: Illustration of variations in α and its consequences for the 2PCF. Top figure: Correlation function of the target halo catalogue (BigMultiDark, crosses) and the grid of $\xi_{\rm H}$ corresponding to the grid of α_7 used for minimisation. The lower sub-panel shows the ratios to the BigMultiDark result. The vertical dashed lines mark the spatial r-range of the fit. Bottom figure: χ^2 (Equation 2.8) as a function of α_7 for the grid of values used in the left panel (red crosses) and the interpolated curve (dashed blue line). In the inner box we zoom into the area near the minimum (green circle).

bin i	$M^i_{\rm th}$ $[h^{-1}M_\odot]$	n_i $[(h^{-1}{\rm Mpc})^{-3}]$	α_i	$f_{\rm vel}$
0	$1.64 \cdot 10^{14}$	$0.05 \cdot 10^{-4}$	4.80	0.564
1	$4.93 \cdot 10^{13}$	$0.45 \cdot 10^{-4}$	2.79	0.672
2	$2.95 \cdot 10^{13}$	$0.95 \cdot 10^{-4}$	2.28	0.715
3	$2.15 \cdot 10^{13}$	$1.45 \cdot 10^{-4}$	2.00	0.743
4	$1.70 \cdot 10^{13}$	$1.95 \cdot 10^{-4}$	1.90	0.754
5	$1.41 \cdot 10^{13}$	$2.45 \cdot 10^{-4}$	1.84	0.760
6	$1.21 \cdot 10^{13}$	$2.95 \cdot 10^{-4}$	1.73	0.771
7	$1.04 \cdot 10^{13}$	$3.50 \cdot 10^{-4}$	1.73	0.771

Table 2.4: Properties of the selected mass bins for the BigMultiDark simulation: mass threshold $M^i_{\rm th}$, equivalent number density $n(M > M^i_{\rm th})$, best fit α_i for the interval of masses $M^{i-1}_{th} < M < M^i_{th}$ and f_{vel} computed for the same interval (see Section 2.4.2).

component velocity distribution of BigMultiDark and the particles selected by halogen. Both curves are well-described by a Gaussian with $\bar{v}_x = 0$, where the standard deviation of the N-body halos is reduced compared to that of $v_{x,\rm p}$, i.e. $\sigma_{\rm p} > \sigma_{\rm NB}$. This confirms our claim in Section 2.2.4 that the particle velocities are larger than the halo velocities, and also shows that a simple linear transformation suffices to map the distribution of $\mathbf{v}_{\rm p} \rightarrow \mathbf{v}_{\rm h}$.

This simple characterisation leads to a transformation of $f_{\rm vel} = \sigma_{\rm NB}/\sigma_{\rm p}$, which is verified by the blue dotted line where this remapping has been applied.

We expect that the velocity bias [138] will be dependent on mass-scale in general. We can easily incorporate this into our fit by calculating

$$f^i_{\rm vel} = \frac{\sigma^i_{\rm NB}}{\sigma^i_{\rm p}} \tag{2.10}$$

for each interval of mass $M = (M^{i-1}_{\rm th} : M^i_{\rm th}]$ while performing the fit for α. These results are also listed in Table 2.4. There is a noticeable decrease in $f_{\rm vel}$ towards higher mass halos. We will see in Section 2.5.4 below how this affects the modelling of Redshift Space Distortions.

We finally note that there may be other more complex models of velocity bias accounting for the physics of low scales and adjusting other statistics beyond the overall

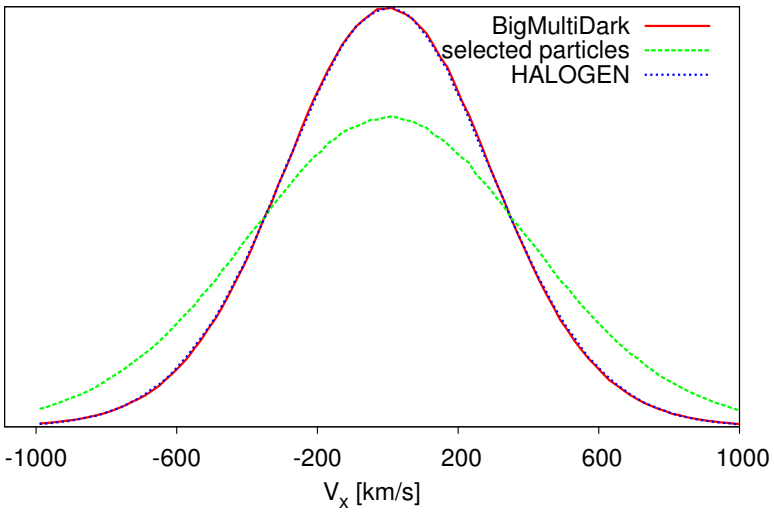

Figure 2.6: One-component (v_x) velocity distribution of the halo catalogues. The FOF halos from the BigMultiDark simulation are in a red solid line, and $v_{x,\mathrm{p}}$ of the particles selected by HALOGEN catalogue are in a green dashed line, while the corrected $\mathbf{v}_\mathrm{h}$ halos from HALOGEN are in a blue dotted line. The correction provides a very closely matching distribution, which has a generally lower velocity.

velocity distribution. However, the model presented here is very simple and capable of reproducing the halo velocity distribution with a great accuracy.

2.4.3 Cell size: l_cell

We have previously mentioned the cell-size l_cell which is introduced to HALOGEN to provide a simple local density via the NGP scheme [137]. We have also noted that it defines a lower-limit of reliability of the resultant 2PCF. In this section we explore this parameter further, describing its effects and how to optimise for it.

In Figure 2.7 we show the 2PCF of the BigMultiDark catalogue against HALOGEN results for several values of l_cell. We note two effects, l_cell

1. determines the minimum scale at which the 2PCF is reliable and

2. controls the broadening of the Baryonic Acoustic Oscillations (BAO) peak.

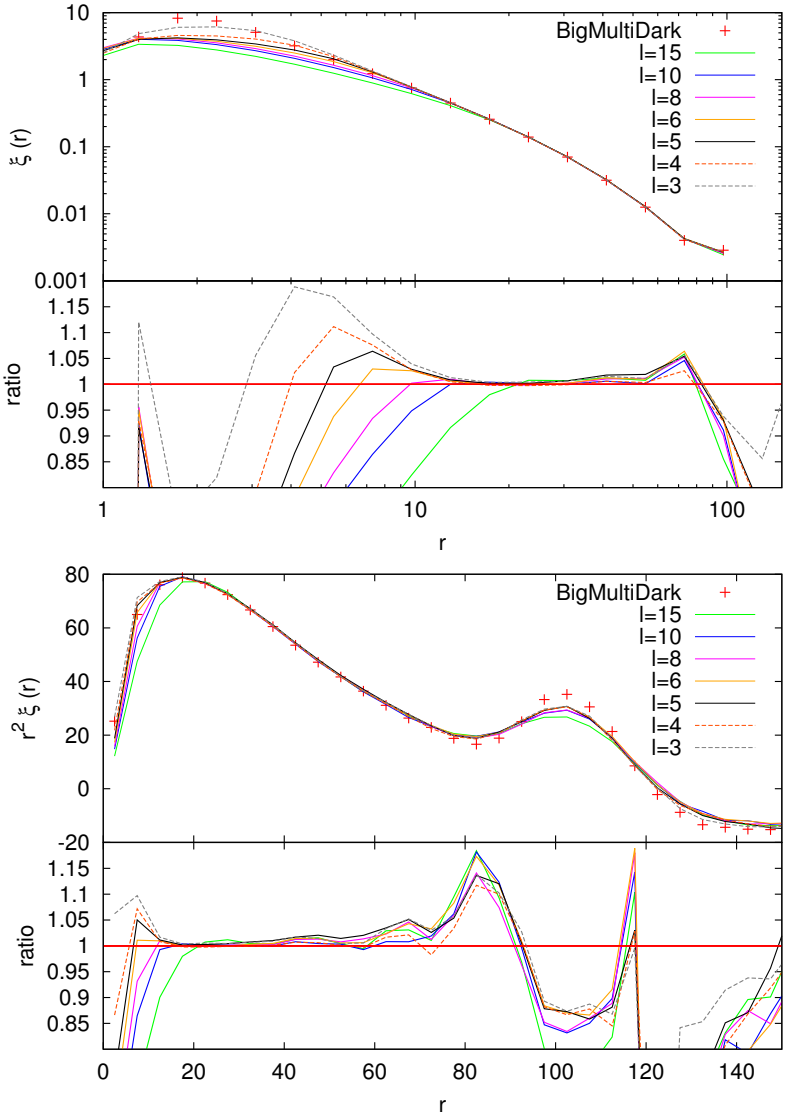

Figure 2.7: Two-point correlation function on logarithmic (top figure) and linear (bottom figure) scale of the FOF catalogue of the BigMultiDark simulation (crosses) against the results from HALOGEN (lines) for different values of l_{cell} (different line styles as indicated in the legend). Note that in the bottom figure the 2PCF has been multiplied by r^2 to increase the visibility of the BAO peak. The lower sub-panels show the ratio with respect to the BigMultiDark curve.

The first effect is clearly noticeable in the top figure where the HALOGEN 2PCF detaches from the BIGMULTIDARK curve at $r \approx l_{\rm cell}$. This is expected, since particles are chosen at random inside the cell, tending towards a bias of unity at these scales.

The second effect is more noticeable in the bottom figure. As $l_{\rm cell}$ is decreased, the broadening and dampening (best seen in the lower sub-panel as the difference between the artificial peak at $r = 80h^{-1}$Mpc and trough at $r = 100h^{-1}$Mpc) is decreased. The reason for this is that we introduce an uncertainty (on a scale $l_{\rm cell}$) in the position of the halos that propagates to an uncertainty in the determination of $r_{\rm BAO}$. In effect, the density field has been filtered by a quasi-top-hat function [137], which has the known effect of peak-broadening.

Clearly, $l_{\rm cell}$ should be set as small as possible to mitigate these effects. However, a limit is enforced by the mean-interparticle-separation, $d_{\rm p}$, of the input density field. We cannot hope to reliably probe scales smaller than $d_{\rm p}$, and even just above this scale we run into the problem of having poor statistics within cells. We recommend using a value of $l_{\rm cell} \geq 1.5 d_{\rm p}$ (ensuring > 3 particles per cell on average), and in this work we take $l_{\rm cell} = 4h^{-1}$Mpc $\approx 2d_{\rm p}$ as the reference.

We comment here that the choice of $l_{\rm cell}$ affects the optimal $\alpha(M)$ relation. This is unfortunate, because it would be useful to be able to perform the fit for α using a lower resolution (since this is the bottleneck). The mechanism by which this effect occurs is known, and we hope to be able to correct for it in the future.

Let us illustrate the mechanism with an example: suppose we take a cell with cell-size $l^{\rm I}_{\rm cell}$ and density $\rho^{\rm I}_{\rm cell}$ from a volume $(N l^{\rm I}_{\rm cell})^3$. For the same distribution, we could also use $l^{\rm II}_{\rm cell} = l^{\rm I}_{\rm cell}/2$, which forms 8 sub-cells i with densities $\rho^{\rm II}_{{\rm cell},i}$. For the same α, the probability of choosing the cell in case I is

$$P^{\rm I}_{\rm cell} = \frac{(\rho^{\rm I}_{\rm cell})^\alpha}{\sum_j^{N^3} (\rho^{\rm I}_j)^\alpha} = \frac{\left(\frac{1}{8}\sum_i^8 \rho^{\rm II}_{{\rm cell},i}\right)^\alpha}{\sum_j^{N^3} (\rho^{\rm I}_j)^\alpha} \tag{2.11}$$

whereas in case II we have

$$P^{\rm II}_{\rm cell} = \frac{\sum_i^8 (\rho^{\rm II}_{{\rm cell},i})^\alpha}{\sum_j^{(2N)^3} (\rho^{\rm II}_j)^\alpha}, \tag{2.12}$$

and clearly these are not in general equivalent if $\alpha \neq 1$. We expect the difference

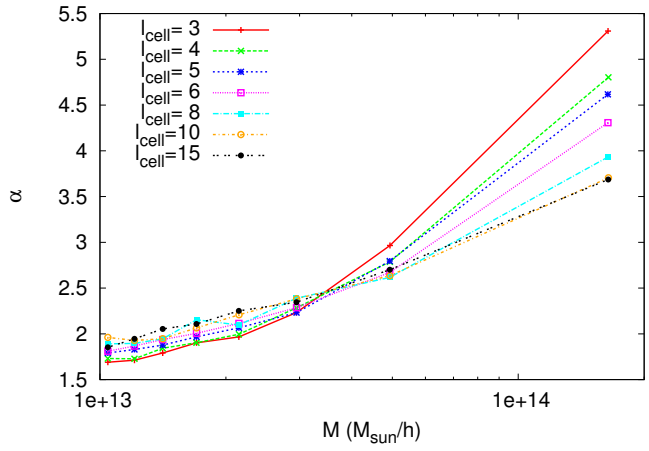

Figure 2.8: Best-fit $\alpha(M)$ functions for different values of $l_{\rm cell}$, as marked in the legend (units of h^{-1}Mpc).

in the distributions to be dependent on α, the two cell-sizes and their ratio and the cosmology, via the mass variance $\sigma(r)$. In future studies we hope to be able to quantify this relationship to enable faster fitting.

Figure 2.8 shows the effect of changing $l_{\rm cell}$ on the best-fit $\alpha(M)$ and we notice two characteristics. Firstly, $\alpha(M)$ is an increasing function for all $l_{\rm cell}$, as expected since $b(M)$ is increasing. Secondly, low masses are less sensitive to $l_{\rm cell}$, which we expect mathematically from Eqs.2.11 and 2.12 with an increasing $\alpha(M)$ (the greater α is, the greater the differences expected).

In Figure 2.7 we have re-fit the $\alpha(M)$ relation for each value of $l_{\rm cell}$, ensuring proper comparison between curves. Furthermore, we run 5 realisations of each and display the average, to reduce the effects of HALOGEN variance.

2.5 HALOGEN: Outcome

While previous sections were dedicated to the design and optimisation of HALOGEN, we have now defined the final method and fixed the optimal parameters. In this section we discuss the performance of HALOGEN in more detail, both in the clustering statistics so far analysed, and in other statistics that HALOGEN is not constrained to

match. We begin by demonstrating the power of HALOGEN for mass-production of halo catalogues for use in deriving covariance matrices to measure cosmic variance, which we envision as the primary application of the HALOGEN machinery.

Some of the results presented here (Section 2.5.2,Section 2.5.3,Section 2.5.4) are also presented in Section 2.6 when comparing to other methods. However, we find some subtleties for which is worth presenting them also here. The PDF (Section 2.5.2) shown here is computed in several cell sizes, exploring different scales. The $P(k)$ in Figure 2.12 is in logarithmic scale, focusing more at large scales. We show in Section 2.5.4 why is necessary to introduce the velocity bias as explained in Section 2.4.2.

2.5.1 Mass production of halo catalogues

The driving motivation of developing fast methods for synthetic halo catalogues is to accurately produce robust covariance matrices for large galaxy survey statistics. Though HALOGEN requires a full N-body simulation to calibrate its two parameter sets, once these parameters have been established, we are free to run as many realisations (with different phases for the initial conditions) of the the halo catalogue (using the same cosmological parameters, volume, mass resolution etc.) as we like. This process is expected to purely simulate the effects of cosmic variance, and thus is extremely valuable for deriving the covariance matrices.

In order to verify that the variance seen in the resulting data traces the expected cosmic variance, we complemented the generation of the HALOGEN catalogues with several corresponding N-body simulations. Due to the computational time constraints, we were only able to run five simulations, which were based on GOLIAT, and in which only the seed for the random Initial Condition (IC) phases was changed. The initial conditions for these runs were generated with 2LPTIC at redshift $z = 32$ (for the N-body) and $z = 0$ (for HALOGEN), using the same seed for each pair. The N-body particle distributions were evolved to $z = 0$ using GADGET2 (and subsequently analysed with AHF).

In Figure 2.9 we present the 2PCF of those 5 pairs of catalogues (random seeds

are colour-coded, with HALOGEN as solid lines, and AHF as points). The HALOGEN lines are the average of 5 realisations of HALOGEN placement (maintaining the same phases) and the error bars show the HALOGEN variance. Given that the GOLIAT box size is rather small ($1h^{-1}$Gpc), scales $r >\sim 60h^{-1}$Mpc are dominated by cosmic variance effects. This makes it easy to identify the signature of each set of initial conditions. Though the realisations are significantly different, we note that the HALOGEN catalogue follows the N-body result, and maintains the correct normalisation at intermediate scales ($20h^{-1}\text{Mpc} < r < 50h^{-1}\text{Mpc}$). We stress that the fitting procedure has only been performed once; all five cases used fixed parameters. The similarity of the goodness of fit in each case (as compared to that directly fitted to) demonstrates that the fitted $\alpha(M)$ is universal with respect to input seed. We note also that the HALOGEN variance is significantly sub-dominant to the cosmic variance.

To better appreciate the dominance of the cosmic variance in a more applicable scenario, we return to the BIGMULTIDARK simulation. This has a reduced cosmic variance due to the larger volume, but has the disadvantage that we cannot run several N-body simulations of this magnitude. The blue line of Figure 2.10 shows how the 2PCF of a single-run HALOGEN (neither HALOGEN nor cosmic variance has been averaged out) compares to the reference BIGMULTIDARK catalogue when they have the same initial condition phases. We further show the HALOGEN variance ($\sigma_{\rm H}$) and cosmic variance ($\sigma_{\rm cosm}$). The former has been computed as usual: running 5 realisations of HALOGEN on the same 2LPT snapshot. For the latter we run five 2LPTIC snapshots with different IC seeds. In order to avoid mixing $\sigma_{\rm cosm}$ and $\sigma_{\rm H}$ for each of them we first averaged out HALOGEN variance by running 5 realisations of HALOGEN and $\sigma_{\rm cosm}$ is computed as the dispersion of the five resulting ($\sigma_{\rm H}$-free) lines. We find for all scales that the HALOGEN variance is dominated by the cosmic variance, $\sigma_H < \sigma_{\rm cosm}$.

2.5.2 Probability Distribution Function

A simple but powerful statistic for point particles is the Probability Distribution Function (PDF), which is the distribution of particles per cell on a given scale.

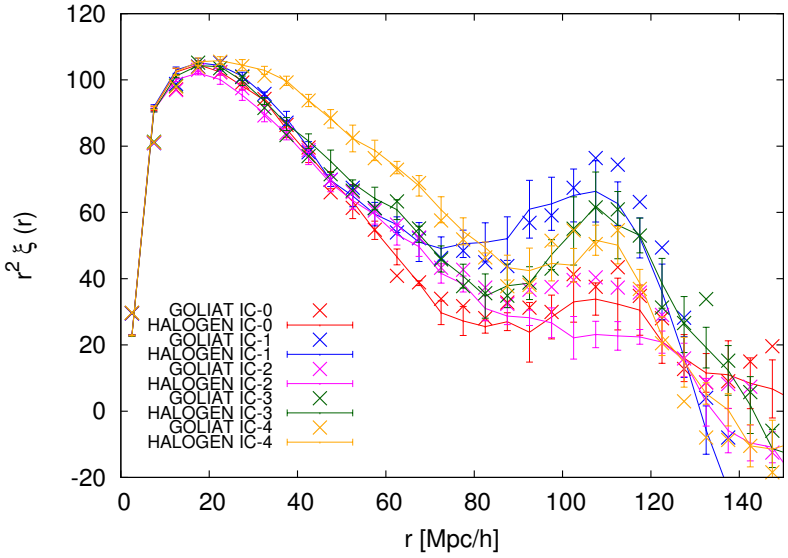

Figure 2.9: 2PCF of HALOGEN (lines) and the AHF (points) catalogues for five different 2LPTic random seeds (colour-coded). The first case corresponds to the original GOLIAT used to obtain the $\alpha(M)$ relation whereas the following share the same setup besides the seed.

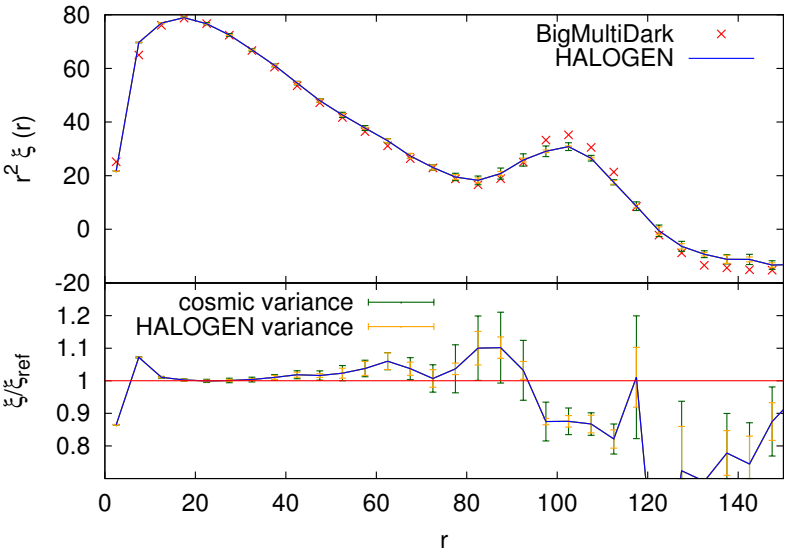

Figure 2.10: 2PCF of the FOF catalogue of the BigMultiDark simulation compared to that of a single-run HALOGEN (non-averaged) with the same initial condition phases. We also include error bars: in green the cosmic variance and in orange the HALOGEN variance (see text). The lower panel shows the ratio with respect to BigMultiDark.

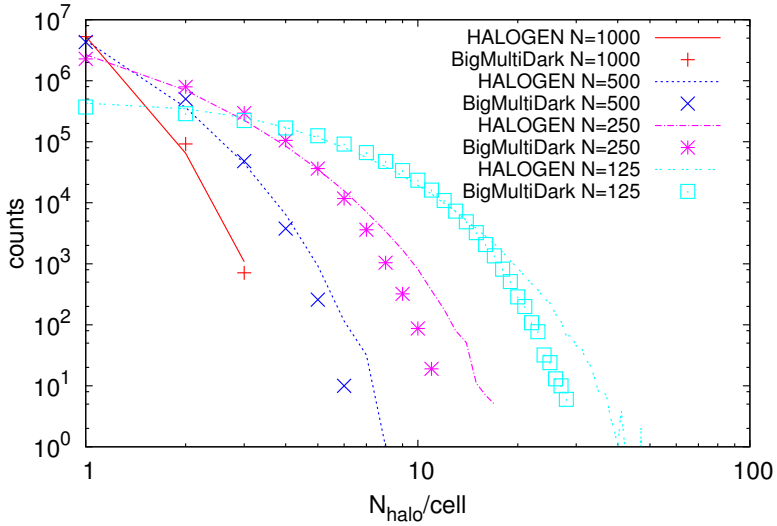

Figure 2.11: PDF of halo counts for both HALOGEN (lines) and BIGMULTIDARK (points) catalogues from BIGMULTIDARK. Several mesh numbers are used, as labelled by colours, and these correspond to the physical scales of $2.5h^{-1}$Mpc, $5h^{-1}$Mpc, $10h^{-1}$Mpc and $20h^{-1}$Mpc respectively.

Though simple, it contains interesting information as it contains contributions from the entire hierarchy of n-point functions [139–141].

Covering the BIGMULTIDARK simulation with meshes of various (regular) sizes, we show in Figure 2.11 a histogram of the number of halos per cell for both the HALOGEN and BIGMULTIDARK catalogues; the cell size ranges from 2.5 to $10h^{-1}$Mpc. We find good agreement, especially at lower numbers of $N_{\rm halo}$/cell, where the contribution of non-linear scales is reduced. We note that the mesh used to calculate the PDF is not to be confused with the grid used by HALOGEN for the NGP density assignment.

2.5.3 Power Spectrum

HALOGEN has been designed to recover the 2PCF $\xi(r)$ of a provided halo catalogue. As the power spectrum $P(k)$ is its Fourier Transform, it theoretically contains the same information. However, this information is distributed differently in the two functions and there is mode coupling when transforming from one to another: an

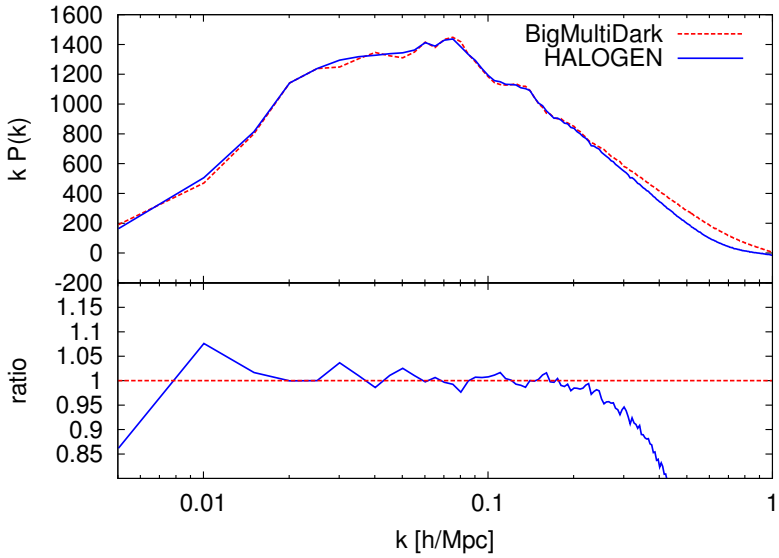

Figure 2.12: Power Spectrum P(k) of HALOGEN (blue line) and FOF (red line) for BIGMULTIDARK. The bottom panel shows their ratio. The Power Spectrum has been computed using a $N = 1024^3$ mesh and corrected for shot noise as explained in [142].

error at a given scale in one of the magnitudes can propagate to an error at all scales in the other. So we expect to witness different strengths and weaknesses in $P(k)$.

In Figure 2.12 we compare the power spectrum of the BIGMULTIDARK FOF catalogue to the corresponding HALOGEN realisation. We find agreement to 5% across the scales $0.01h\text{Mpc}^{-1} < k < 0.3h\text{Mpc}^{-1}$, but note that smaller scales $k > 0.3h\text{Mpc}^{-1}$ ($r < 20h^{-1}\text{Mpc}$) are underestimated. This underestimation arises from the smallest scales of the 2PCF, $r < l_{\text{cell}}$, which integrate through higher scales in $P(k)$.

2.5.4 Correlation Function in Redshift Space

Observed galaxies are not directly located in 3D space, but 2D-angular (θ, ϕ) with redshift z converted to a polar distance. However, such distances are modified by galaxies' peculiar velocities – velocity components that are not due to the Hubble expansion. These modifications are encoded as Redshift Space Distortions (RSD), and we can begin to account for them by assigning correct velocities to halos.

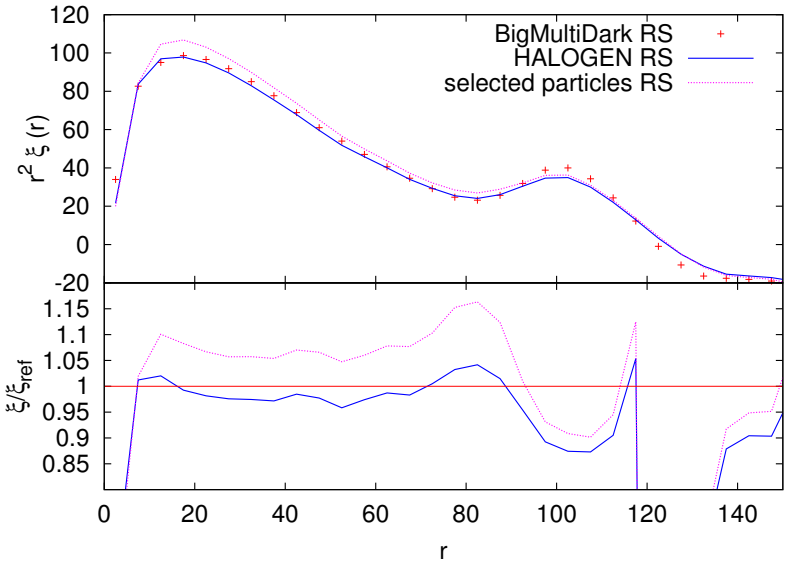

Figure 2.13: 2PCF in redshift space (RS) for FOF (red points), and HALOGEN (blue line) of the BIGMULTIDARK simulation. We also include in magenta the results of our catalogue without applying the velocity bias (i.e. $f_{\rm vel} = 1$, 'selected particles') and find that a correct velocity bias is needed.

Using the halo velocities, we can mimic this effect when calculating the 2PCF. We show the results of such an analysis in Figure 2.13, in which the monopole of the 2PCF in redshift space is compared for the HALOGEN and BIGMULTIDARK catalogues. To show the effect of our velocity transformation, we also include the 2PCF of the 'selected particles' in which the velocities were not transformed. The normalisation and shape are significantly improved by the simple linear transformation (Equation 2.9), and we find agreement to below 5% per cent at intermediate scales.

2.6 Comparison with other Approximate Methods

So far, we devoted this chapter to the construction and analysis of HALOGEN. However, there are other approximate methods in the literature that also generate fast halo mock catalogs. Within the Mocking Astrophysics program described in Sec-

	BigMD (NB)	COLA	EZmock	HALOGEN	Log-normal	PATCHY	PINOCCHIO	PTHalos
CPU-hour	800,000	130	1.3	6.7	0.5	8	440	45
Memory	8Tb	550Gb	28Gb	130Gb	15Gb	24Gb	890Gb	112Gb
Particle (force) mesh size	3840^3	1280^3 (3840^3)	960^3	1280^3	1280^3	960^3	1920^3	1280^3
Resolve halos	YES	YES	NO	NO	NO	NO	YES	YES

Table 2.5: Computing resources and related properties used by each code to generate the halo catalogue analysed in this study. The CPU-hours can vary significantly from one machine to another, but it is important to note their order of magnitude, which depends on the algorithm and the particle mesh size. The memory usage is mostly determined by the mesh size, that determines the spatial and mass resolution. Whereas most codes use the same particle and force mesh, COLA need 3 times more resolution in the latter. Codes that need to resolve halos need more particles and, hence, more resources, but always much lower than a full N-Body simulation (BigMD).

tion 1.1.3, the 'nIFTy cosmology' workshop[14] arose, in which we compared nearly all existing methods for approximate halo mock catalogs.

We briefly present here some of the results that emerged from that comparison [3].

2.6.1 Description of methods

As described before (Section 2.1.1 & Section 2.2), all methods can be seen as a four step process. The main differences among methods rest in the way they generate the density field and how they apply a bias to generate a halo distribution. This idea is graphically represented in Figure 2.14.

The methods presented here can be used in different contexts and each of them is designed with different purposes. Some of them require less computing resources at the price of having lower resolution, whereas others prefer to keep the resources higher but gain accuracy. Table 2.5 compares the resources needed for each method to generate the same halo catalogue (being the N-Body simulation BigMultiDark in Table 2.1 the reference catalogue). Those that need to resolve halos (COLA, PINOCCHIO and PTHalos) have a predictive nature and typically require more resources than those with a stochastic nature (EZmocks, HALOGEN, PATCHY and Log-Normal) that need to be fitted to a reference simulation.

[14] http://popia.ft.uam.es/nIFTyCosmology

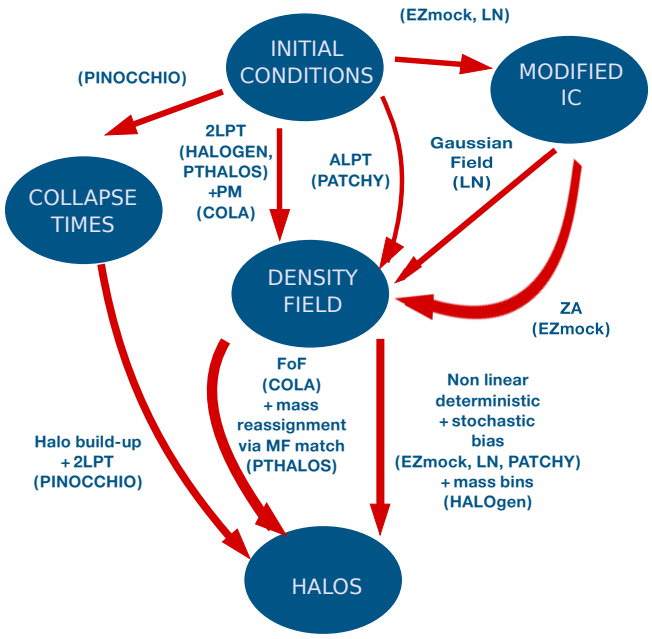

Figure 2.14: Scheme of approximate methods. Most of them use a gravity solver (2LPT, ALPT, 2LPT+PM, ZA) to generate a density field from which halos are generated either using a halo finder or a stochastic bias. Some methods additionally need to modify the initial power spectrum (EZMOCKS and LOG-NORMAL). PINOCCHIO computes the halo formation and evolution in collapse time.

The main characteristics of the methods are shown in Table 2.6, and we briefly describe them hereunder:

- COLA [COmoving Lagrangian Acceleration, 119] is a PM method (Section 1.1.1) in which the equations of motion have been rewritten by subtracting the 2LPT solution $\vec{x}_{\rm res} = \vec{x} - \vec{x}_{\rm 2LPT}$. Then $\vec{x}_{\rm 2LPT}$ is computed following LPT (as explained in Section 1.1.1), and $\vec{x}_{\rm res}$ can be integrated in larger timesteps, saving substantial computational time. The halos are extracted with a halo finder.

- EZMOCKS. [Effective Zel'dovich approximation mock catalogue, 121] is constructed from the Zel'dovich approximation density field. It is an stochastic method that maps the PDF from the reference simulation and fits several parameters (density saturation, density threshold, $P(k)$-tilt, BAO-enhancement, etc.) to obtain the correct 2-point and 3-point functions.

- HALOGEN. [2], extensively described in this chapter, is a stochastic method that fits the large scale bias with one parameter (power-law bias), by placing halos in a 2LPT density field.

- LogNormal. At large scale, galaxies have been measured to follow a lognormal distribution [143, 144], and this can be derived from the continuity equation in linear perturbation theory if the initial conditions are gaussian [114]. This method places halos following a lognormal distribution that matches by construction any desired correlation function (up to certain scales), however it lacks any physics of any higher order.

- PATCHY[118] solves gravity with Augmented-LPT ([145]), a combination of 2LPT at large scales and spherical collapse model at small scales. It uses a non-linear, scale-dependant and stochastic biasing prescription based on several parameters (density threshold, density cut-off, power-law, etc.) fitted to match the PDF and power spectrum.

- PINOCCHIO[116] is based on the ellipsoidal collapse, solved with the aid of 3LPT to compute the time at which mass elements collapse (in the orbit-crossing sense), and Extended Press & Schecther (EPS) to deal with multiple smoothing radii. It starts from the generation of a a regular grid in Lagrangian space, the density field is smoothed on a set of scales, and the collapse time is computed for each particle and at each smoothing radius. The earliest time is recorded as the estimate of collapse time. An algorithm mimics the hierarchical formation and merging of halos, and collapsed objects are moved with 2LPT, finally generating both the halo catalogue and merger tree.

- PTHALOS[115, Perturbation Theory halos] is based on a 2LPT density field from which halos are found using a friends-of-friend algorithm. Since matter collapses differently in 2LPT than in an N-Body simulation the linking length used is $b_{\rm 2LPT} = b_{sim} \left(\frac{\Delta_{vir}^{sim}}{\Delta_{vir}^{2LPT}} \right)^{(1/3)}$.

2.6.2 Results

Here we study how different methods perform in the 1, 2 and 3-point statistics. Recall that the PDF is primarily a 1-point statistics, but with contributions of all higher orders. This section focuses more in the 2-point functions, which are the most studied in the literature (and the primary objective of HALOGEN) because it contains the most net information about cosmology (including BAO). 3-point functions are more difficult to measure with current surveys (although it has been done [147]) and have high contributions from non-linearities more difficult to predict from the theory. Nevertheless, it is also a target for the future surveys.

The PDF distribution is shown in Figure 2.15, where we find two outliers: the LOG-NORMAL method and PINOCCHIO. Note however, that the scales explored here $2.6 Mpc/h$ are already highly non-linear.

In regard to the 2-point function, looking at the ξ at the top part of Figure 2.16, we find that most methods give similar results both in real (left) and redshift (right) space. For the LOG-NORMAL method, velocities where not computed (although they can be computed with linear theory), so all results from redshift space are missing. The normalisation of PTHALOS is off by more than 20%, this is due to the fact that here we took the binding length b_{2LPT} from a theoretical value and categorised PTHALOS as a predictive method, but b_{2LPT} could be left free and the bias fitted.

In Fourier space (bottom of Figure 2.16), it appears similarly at the linear scales, but this figure focuses more in the non-linear scales ($k > 0.1 h/Mpc$, compared to Figure 2.12) where methods based on 2LPT (HALOGEN, PTHALOS and PINOCCHIO) and LOG-NORMAL start having problems. Only methods with accurate density field (COLA and PATCHY) or many free bias parameters (EZMOCKS) can reproduce these scales within 5% error.

For the 3-point function something similar occurs: we need more sophisticated methods. Particularly, the LOG-NORMAL does not reproduce even the shape of the functions, whereas HALOGEN and PTHALOS (and slightly PINOCCHIO) have an offset in the normalisation but reproduce the shape.

In conclusion, as long as the 2-point is concerned, nearly all the methods presented

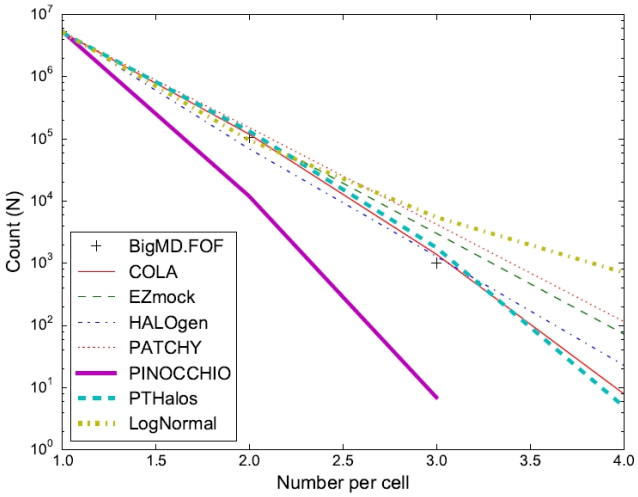

Figure 2.15: PDF of halo counts in a grid with $N = 960^3$ cells for the different approximate methods.

here can reproduce good results at large scales. This is particularly interesting for BAO analysis. If we are also interested in higher order statistics we will need more sophisticated methods that may require more computing resources or a more complex bias model. Depending of the needs of a particular study it will be more convenient to use one code or other.

2.7 Conclusions

We have presented a method called HALOGEN for the construction of approximate halo catalogues. It consists of 4 major steps:

1. Create a distribution of particles in a cosmological volume using 2^{nd}-order Lagrangian Perturbation Theory and distribute them in a grid of cell size l_{cell}
2. Sample a theoretical halo mass function $n(> M)$ with a list of N_{h} halo masses M and order them in descending mass.
3. Place the halos at the position of particles with a probability dependent on

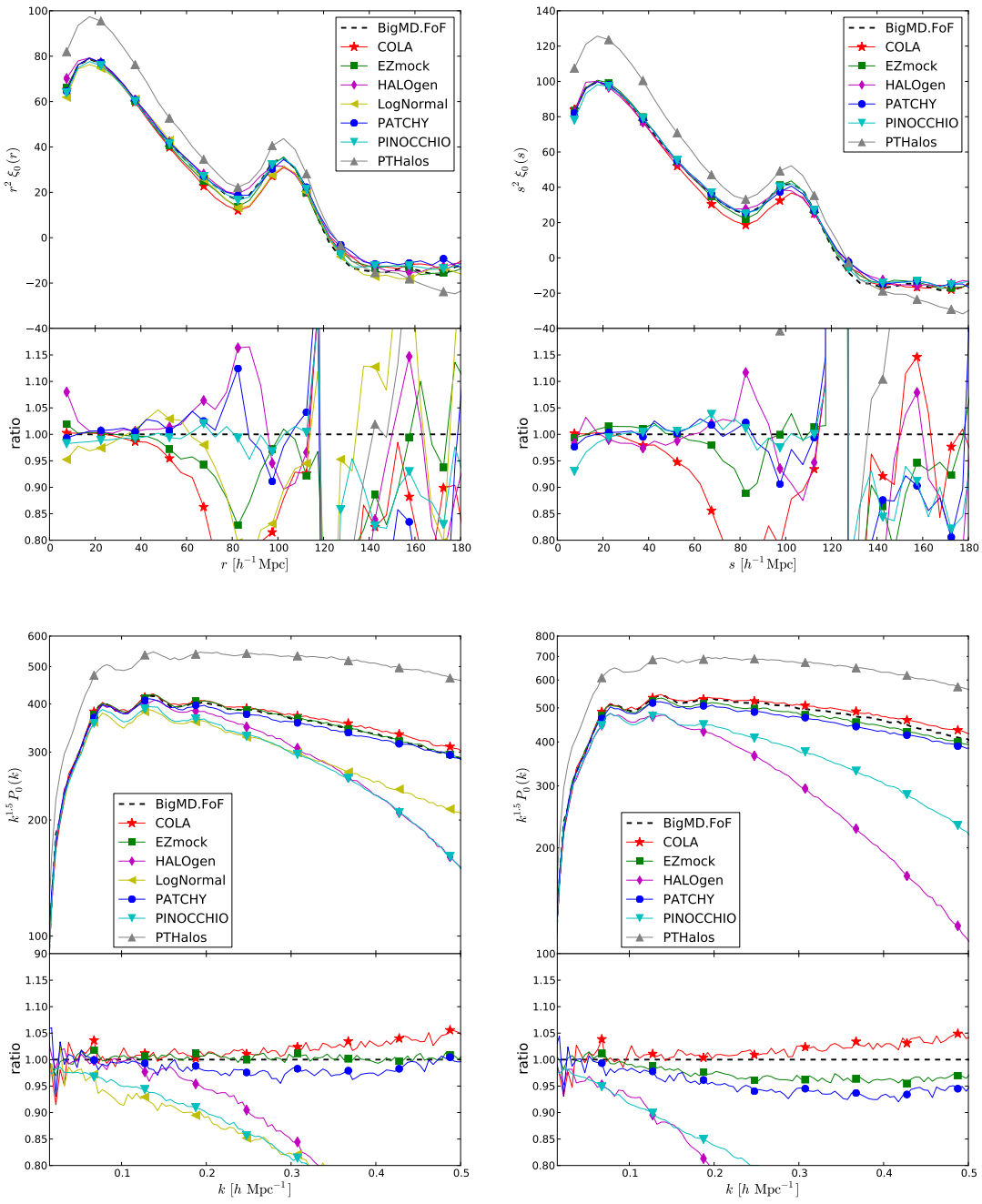

Figure 2.16: Comparison of the 2-point functions for different methods. Top sub-figures show configuration space whereas bottom panels show Fourier space. Left sub-figures show real space and right subfigures are represented in redshift space.

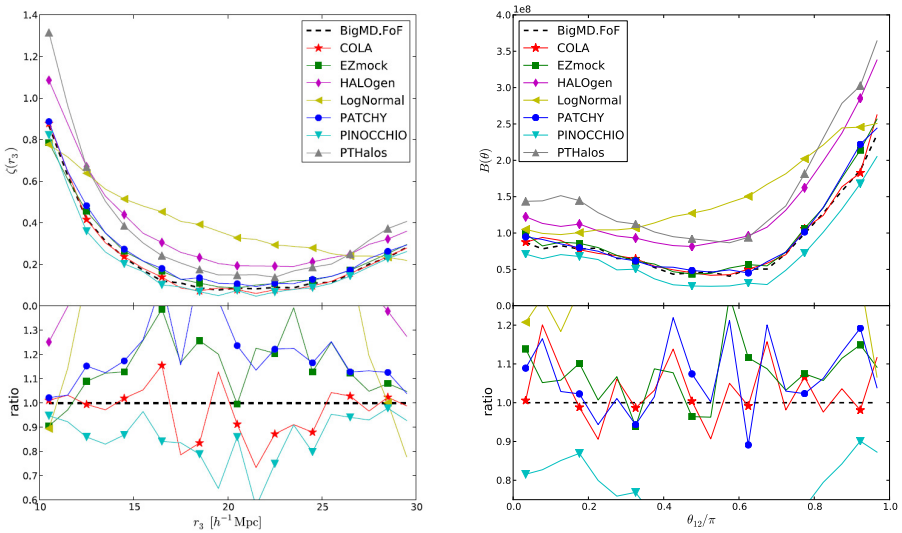

Figure 2.17: Comparison of the 3-point functions for different methods. Left: 3-point function in real space with fixed $r_1 = 10h^{-1}$Mpc and $r_2 = 20h^{-1}$Mpc and free r_3 Right: Bispectrum with $k_1 = 0.1\,h\,\mathrm{Mpc}^{-1}$ and $k_2 = 0.2\,h\,\mathrm{Mpc}^{-1}$, and a varying angle θ_{12}.

the cell density and halo mass $P_{\mathrm{cell}} \propto \rho_{\mathrm{cell}}^{\alpha(M)}$. We select random particles within cells, respecting the *exclusion* criterion and conserving mass in cells (cf. Section 2.3).

4. Assign the velocity of the selected particle to the halo through a factor $\mathbf{v}_{\mathrm{halo}} = f_{\mathrm{vel}}(M) \cdot \mathbf{v}_{\mathrm{part}}$

Further, we noted the modularity of these steps and acknowledged alternatives for each of them. The 2LPT in step (1) provides us with the correct large scale clustering at a low computational cost, while step (2) reconstructs the halo mass function. The heart of HALOGEN is step (3) where the mass dependent bias is modelled through the parameter $\alpha(M)$ that stochastically places more massive halos in overdensities, recovering the correct 2-point correlation function as a function of mass. We also preclude halos from overlapping to match the small-scale behaviour of the 2-point clustering. In the last step (4), we re-map particle velocities in order to obtain the correct halo velocity distribution.

We studied how the parameters of the method – $\alpha(M)$, $f_{\rm vel}(M)$ and $l_{\rm cell}$– can be optimised and summarised the results in Table 2.3. Though HALOGEN needs a reference halo catalogue from an N-Body simulation to obtain $\alpha(M)$ and $f_{\rm vel}(M)$, once they have been optimised for a given setup, HALOGEN can be used to generate a multitude of halo catalogues, allowing the quantification of cosmic variance.

The halo mass function is recovered by construction to the theoretical value. The 2-point function at intermediate scales ($10h^{-1}$Mpc $< r < 50h^{-1}$Mpc, where the bias is controlled by $\alpha(M)$) can be obtained in a BIGMULTIDARK-like simulation at the $\sim 2\%$ level and to the 15% level at BAO scales ($80h^{-1}$Mpc $< r < 110h^{-1}$Mpc) (Figure 2.10). In redshift space, the error at intermediate scales rises to $\sim 4\%$ and remains at $\sim 15\%$ at large scales (Figure 2.13). The clustering has a mass-dependence, for which the accuracy is controlled by the number of bins in the $\alpha(M)$ fit (Figure 2.4). The power spectrum can be recovered at the 5% level in the range of scales $0.01{\rm Mpc}^{-1}h < k < 0.3{\rm Mpc}^{-1}h$ (Figure 2.12). The halo PDF is accurately reproduced at low N_{halo}/cell, but overpredicts the high-$N_{\rm halo}$/cell tail where the contributions of non-linearities are higher (Figure 2.11).

HALOGEN was constructed in favour of simplicity of the method and adaptability. Even though GOLIAT and BIGMULTIDARK have different characteristics (see Table 2.1), HALOGEN can be used for both with little recalibration effort. In Section 3.2.1 we will also fit it to MICE simulation, with still another very different setup. This indicates that HALOGEN is not only capable of running on one specific box-size, redshift or cosmology, which makes it a powerful tool for exploring the statistics of varying cosmologies etc.

We have also verified that changing the initial phases in 2LPTIC for HALOGEN leads to changes in the correlation function (due to cosmic variance) that follow the N-body simulation both in shape and normalisation. This implies that doing so will yield robust estimates of cosmic variance, over potentially hundreds to thousands of realisations. Hence, it has been demonstrated that HALOGEN is a powerful tool for modelling statistics of halo catalogues, and quantify the effects of cosmic variance on them.

Comparing HALOGEN with other methods, we find that the 2-point correlation function at large scales is well recovered by nearly all methods including HALOGEN. HALOGEN is also found well suited for PDF statistics. If we also want to recover non-linear scales or 3-point functions, a more sophisticated method would be required. This method could either have a very accurate density field as COLA, for which computing resources are high compared with a statistical approach as HALOGEN, or a complex bias model with many free parameters that need be tuned to recover all the different statistics (as PATCHY and EZMOCKS), losing in adaptability and simplicity. This links with the idea of modularity remarked across the chapter, we could change the density field (step 1) or the way we place halos (step 3).

For example, for BAO physics, where only large scales of the 2-point function are relevant or for Counts-in-Cells (observational counterpart to PDF), HALOGEN has been demonstrated to be a powerful tool able to generate fast mock catalogues, with low computing resources and simple algorithms. In the next chapter we will see an example of exactly this: how HALOGEN is used to study the systematics and account for the cosmic variance of an experiment, and show how eventually will be used to determine the error bars of a BAO measurement.

Approximate halo mock generation is an emerging field that will have great impact in the coming years with the increasing volume surveyed by the experiments. For different studies there will be a different optimal method depending on the accuracy needed, computing resources available, adaptability to different cosmologies required, number of catalogs needed, etc. Having a variety of methods available and knowing the strengths and weaknesses of all of them will be crucial for the experiments. The new era of observational cosmology is moving forward fast and cosmology modelling must adapt its pace for the new times.

	COLA	EZMOCKS	HALOGEN	LOG-NORMAL	PATCHY	PINOCCHIO	PTHALOS
Mass, Vel	M + V	M(post-process) + V	M(binned) + V	–	M(post-process) + V	M+V	M +V
Initial conditions	2LPT	ZA	2LPT	Gaussian	ALPT	2LPT	2LPT
Using white noise	NO	YES	YES	NO	YES	YES	NO
Assumed HMF	NO	YES	YES	–	YES	NO	YES
Assumed bias model	NO	YES	YES	NO	YES	NO	NO
Substructures	Post-process	YES	Post-process	Yes	YES	Post-process	Post-process
Merger histories	NO	NO	NO	NO	NO	YES	NO
No. free params	0	7	1 (each mass bin)	–	7	5	1
No. free params for z-space dist.	0	1	1	–	2	0	0
No. free params for HMF	0	–	adopt HMF	–	–	5	adopt HMF
No. free params for bias	0	6	1	–	5	0	0

Table 2.6: Main technical features of the methodologies. From top to bottom: whether they provide mass and velocity, how they generate initial conditions, whether they used the same initial random seeds as BIGMULTIDARK (COLA did not and large scales could be affected by cosmic variance), whether they generate or assume a halo mass function (EZMOCKS and PATCHY generate it with a post-processing procedure explained in [146]), whether they assume a bias model, whether provide substructure and merger trees, the number of free parameters introduced in total, for the RSD, for HMF and for the bias.

Chapter 3

Dark Energy Survey Galaxy Mock Catalogues

3.1 The Dark Energy Survey

The Dark energy Survey (DES) [19] is a photometric survey designed to observe the southern hemisphere sky. In particular, DES aims at constraining the equation of state $w(a)$ of Dark Energy in order to shed light on its nature. For that it combines four different main probes:

- Baryon Acoustic Oscillation (BAO)
- Type Ia Supernova (SNIa)
- Galaxy Cluster Counts
- Weak Lensing (WL)

Observations are performed with the 570-Megapixel digital Dark Energy camera (DECam) mounted on the 4-meter Victor Blanco Telescope in Chile. DECam was specifically designed for this experiment. Its main peculiarity is its high sensitivity at the red end of the visible spectrum and at the near infrared, crucial for the detection of objects at high redshift. The survey will cover 5000 $\deg^2$ using a field of view of

2.2 deg of diameter with five different filters (the traditional g, r, i, z and the infrared Y) over 5 years, reaching a magnitude limit of 24 in the band i. DES will observe ~ 200 million galaxies up to $z \sim 1.4$ determining its angular position, photometric redshift (photo-z) and shape.

Opposed to spectroscopic redshift surveys, where redshift can be measured accurately ($\sigma_z \sim (0.001 - 0.0001)(1 + z)$) with a spectrograph, DES is a photometric survey where redshifts are estimated by the combination of flux obtained in the 5 filters (see some techniques in [148–150]) with a typical accuracy of $\sigma_z \approx 0.03(1 + z)$. This decreases the knowledge we have about galaxy *radial positions*, but at the same time it allows us to increase significantly the number of observed galaxies (as spectra measurements are very time consuming), obtaining a complete magnitude-limit survey. Additionally, problems with fibre collisions and apertures, associated with spectroscopy, do not appear here.

Baryonic Acoustic Oscillations (BAO) of the primordial photon-matter plasma leave their imprint in the large scale distribution of matter. From galaxy positions, we can measure the correlation function and find the BAO feature. Detecting BAO with a photometric survey will be an arduous task, since most of the radial information is lost, and the density field is effectively smoothed. But, carefully analysing the data, DES will be able to detect the BAO scale evolution with redshift in the range $0.6 \lesssim z \lesssim 1.4$ and, consequently, measure the evolution of expansion of the Universe. More importantly, this is a range not explored before with BAO, and it will tighten the constraints in the distance-redshift relation shown in Figure 2.

Galaxy shapes are distorted due to gravitational lensing. Whereas in some cases, this effect is so strong that we see multiple images of the same galaxy (strong lensing), generally is much milder and can not be seen in individual galaxies (as intrinsic dispersion of shapes is larger), but only study it statistically. This phenomenon is known as Weak Lensing and tells us about the amount and clustering stage of dark matter. Galaxy cluster counts is another mean to measure the dark matter and its stage of clustering, as is tightly related to the high mass halo abundance as seen in simulations. In the standard ΛCDM model we expect to detect over $100,000$ clusters with DES (being sensitive to clusters with ~ 10 red-sequence galaxies). Studying these two effects as a function of time will be another probe of the expansion of the

universe.

SNIa are used as standard candles in Cosmology to study the evolution of the Universe and were the first evidence of Dark Energy. DES has 4 special fields for SNIa search different to the galaxy field (Figure 3.1), as for SNIa we need to target the same field periodically to search new appearing objects and characterise their light curves (flux as a function of time). Each DES supernova field is revisited ~ 5 times every month, and will discover ~ 4000 SNIa up to redshift $z \sim 1$.

All probes combined together will tightly constrain the time-dependent equation of state of Dark Energy parametrised as $w = w_0 + (a-1)w_a$, as already indicated at the bottom of Figure 1. But DES is well suited for many more astrophysical studies. From DES early data, there have been many remarkable discoveries [151]: 17 (out of 48 known) Milky Way satellite galaxies, a new type of objects termed Super-luminous SN, high-redshift ($z \sim 6$) and lensed quasars, 34 new transnewtonian objects, etc.

DES is also relevant for the two major astrophysical events that happened in the last few months and that even reached the public attention: the discovery of a ninth planet in the Solar System [152] and the direct detection of gravitational waves by the LIGO experiment [153]. DES has an agreement with LIGO to search for an optical counterpart of any triggered detection of gravitational waves. No optical counterpart was found for LIGO event GW150914 [154, 155], caused by a merger of two massive black holes. This is not surprising, since this type of merger is not expected to emit in the optical, but it can be really useful for other type of events or to find unexpected physics. As for the ninth planet, the predicted trajectory [156] passes through the DES observed area, so that a detection may be possible in the future.

The DES data are split by seasons into Science Verification (SV), Year-1 (Y1), Year-2 (Y2), etc. The Science Verification observations were taken in 2012 and 2013 and provided data of over 250 deg^2 at nearly the nominal depth of DES. It was used to test all the science potential of the 5-year survey, finding promising results for cosmology [157–162]. While SV has been widely analysed, DES collaboration is now analysing Y1 post-processed data taken between August 2013 and February 2014. Y1 covers a large fraction of the targeted area but at a milder flux limit (Figure 3.1). Y2

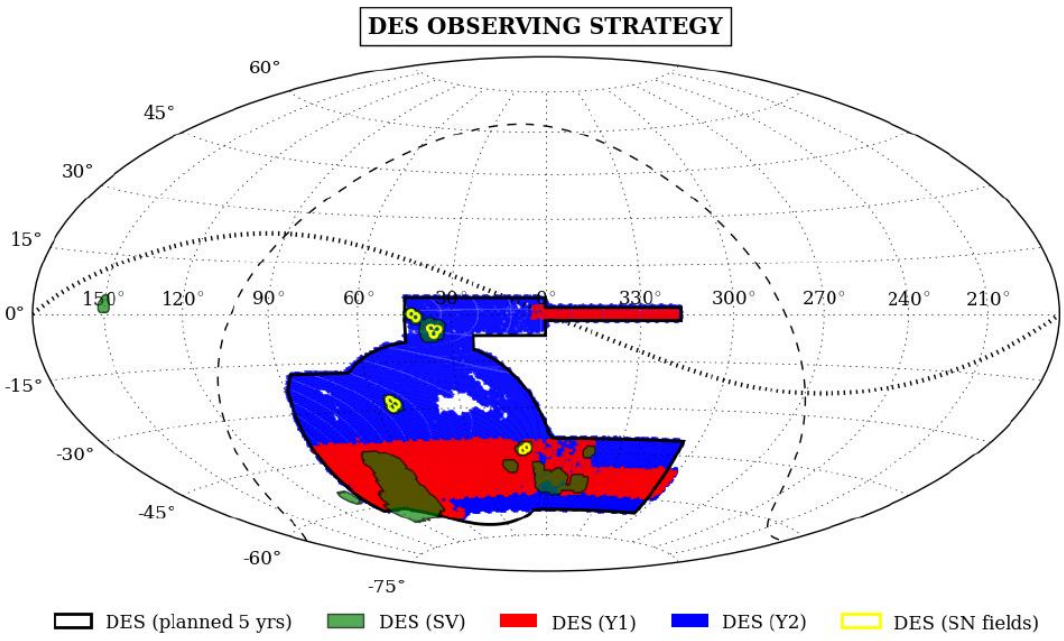

Figure 3.1: DES observed strategy footprint from [151]. We find the supernova field, SV regions, and Y1, Y2 and Y5 masks in equatorial coordinates. Dash and dotted lines represent, respectively, the galactic and ecliptic planes.

is currently being post-processed, although some results quoted above have already emerged from it.

This chapter is part of the work done within the Large Scale Structure Working Group with the aim of detecting BAO from Y1 data. It focuses particularly in the creation of galaxy mock catalogues matching the overall statistics of the selected LSS-Y1 sample (see Section 3.2) to be used to compute the covariance matrices and error bars on the large scale clustering. We further present preliminary results of its application for the data analysis (Section 3.3).

3.2 HALOGEN lamps: observational galaxy mock catalogues

In Chapter 2 we described the HALOGEN method to generate halo mock catalogues in a simulation box, i.e. a distribution of halos in cartesian comoving coordinates at a fixed cosmological time. However, these dark matter halos are not direct observables, we need to include their luminous counterpart: galaxies. Here, we present HALOGEN-LAMPS, a new implementation of HALOGEN with three new observational features:

- **Lightcone**. From observations we do not measure cartesian coordinates (X,Y,Z) at a fixed time t but angular positions and redshift –a combination of radial, temporal and velocity information– (ra, dec, $z_{\rm rsd}$)[1]. This effect is included in Section 3.2.1
- **Photo-z**. DES is a photometric survey for which $z_{\rm rsd}$ is estimated by $z_{\rm ph}$ with low precision. This effect mixes galaxies of different z in the same $z_{\rm ph}$-bin, we will see how to implement it in Section 3.2.2.
- **Galaxy population**. We do not observe dark matter halos but galaxies, which show different clustering. We generate galaxy catalogues with a HAM and HOD method in Section 3.2.3.

In this section we simulate these three effects with the aim of creating galaxy mock catalogues with the same statistical properties as the selected LSS-Y1 sample. Namely, the same galaxy number density as a function of redshift $n(z_{\rm ph})$, the same angular correlation function in $z_{\rm ph}$-bins $w^i(\theta)$ and the same $P(z_{\rm rsd}|z_{\rm ph})$ distribution.

The selection of the LSS-Y1 sample has been optimised to yield a BAO detection with error below the 5%. It consists in a sub-sample of the full Y1 data, to which we apply three main cuts in the different filter magnitudes: completeness $17.5 < m_i < 22$, brightness $m_i < 19 + 3z_{\rm ph}$ and red selection $(m_i - m_z) + 2(m_r - m_i) < 1.7$. The

[1]From now we will use $z_{\rm rsd}$ as the ideally measured redshift with no error but with redshift space distortions included. We introduce this notation to distinguish it from the $z = z_{\rm true}$ that represents the cosmological time and has been used so far, and also from $z_{\rm ph}$.

selection has been done balancing the trade-off between having a sample with higher bias and better photo-z (brighter and redder sample) and reducing the shot-noise (that increases if we reduce the number density). This is optimised together with the selection of the mask, based on the goodness of the different areas (see [163] for the details).

This links with a forth observational feature: the application of a mask. A mask consists in a list of pixels (we use healpix pixelisation[2]) telling us which regions of the sky can be used, and which ones can not. Excluded regions can be due to no observation, insufficient observation (for magnitude limited samples), bad seeing, foreground (mainly stellar) contamination or other causes for large systematics. This leads to a somewhat patchy footprint that depends specifically on the selected sample. The red region in Figure 3.1 represents approximately the Y1 mask. More specifically, the Y1-LSS mask has an area of $\sim 1426 deg^2$. This mask does not fit in an octant (it is around 150° wide in ra), and we need to cover it with three different patches of the lightcone generated in Section 3.2.1. We will not enter into the details of masking, beyond noting that the Y1-LSS mask has been applied to all the catalogues analysed in the figures from Figure 3.4 onwards.

3.2.1 Lightcone

Changing from cartesian coordinates to an observational lightcone is very simple once the observer is placed. We place the observer at one corner of the box, so that we can simulate one octant of the sky, and transform coordinates as

$$\begin{aligned} \text{ra} &= \arctan\left(\frac{Y}{X}\right) \\ \text{dec} &= \arcsin\left(\frac{Z}{r}\right) \\ z_{\text{rsd}} &= z(r) + \frac{1+z(r)}{c}\vec{u}\cdot\hat{r} \end{aligned} \tag{3.1}$$

being $r = \sqrt{X^2+Y^2+Z^2}$, $\vec{u}$ the comoving velocity, $\hat{r} = \vec{r}/r$ and $z(r)$ the inverse of

[2] http://healpix.sourceforge.net/

$$r(z) = c \int_0^z \frac{dz'}{H(z')} \tag{3.2}$$

The reason why this implementation is called a lightcone is because the cosmological time ($t(z)$) is determined by the radial distance r in the same way as done in observations as light travels. But the Universe changes with time and hence, our simulations will also change with redshift.

Particularly, we are interested in a redshift-dependent clustering, and hence we will have the HALOGEN parameters (α and $f_{\rm vel}$, summarised in Table 2.3) varying as a function of redshift as well. For this we will use as a reference the MICE N-Body simulation (see Table 2.1 and Section 2.1.2) and fit $\alpha(M)$ and $f_{\rm vel}(M)$ at the snapshots $z = 0, 0.5, 1.0, 1.5$ and interpolate at intermediate redshifts.

The outcome of the fit is shown in Figure 3.2 where we see the mass-dependent clustering for the snapshots $z = 0.5$ and $z = 1.0$. Note, that the reference density for this simulation is ~ 4.5 times bigger than the one previously used for BIGMULTIDARK, and the minimum mass $M_{\rm min}$ here 4 times smaller. This is roughly the minimum number density that we need to simulate the sample. We found that in this case a logarithmic binning of masses was more useful, and we represent in Figure 3.2 the Mass thresholds that were used during the fitting.

The HALOGEN parameters (including the HMF) were interpolated to $z = 0.55$, 0.625, 0.675, 0.725, 0.775, 0.825, 0.875, 0.925, 0.975, 1.05 and HALOGEN was run at those redshifts. We build the lightcone from the superposition of $z_{\rm rsd}$ shells of those snapshots by setting the edges at the intermediate redshifts, and saving data from $0.45 < z_{\rm rsd} < 1.2$ (restricted for storage saving). We repeat the same process 8 times setting the observer in the 8 corners of the box to generate 8 different catalogues. This process might not be ideal and we are working on a future version of the catalogues where we avoid the need for 10 snapshots by building the lightcone directly in one box with growth factors that depends on the position $D_{1,2}(z(r))$ in the 2LPT Equation 1.5.

Finally, we compare the resulting HALOGEN lightcone with the halo lightcone generated by MICE in Figure 3.3. The MICE simulated lightcone is constructed from fine

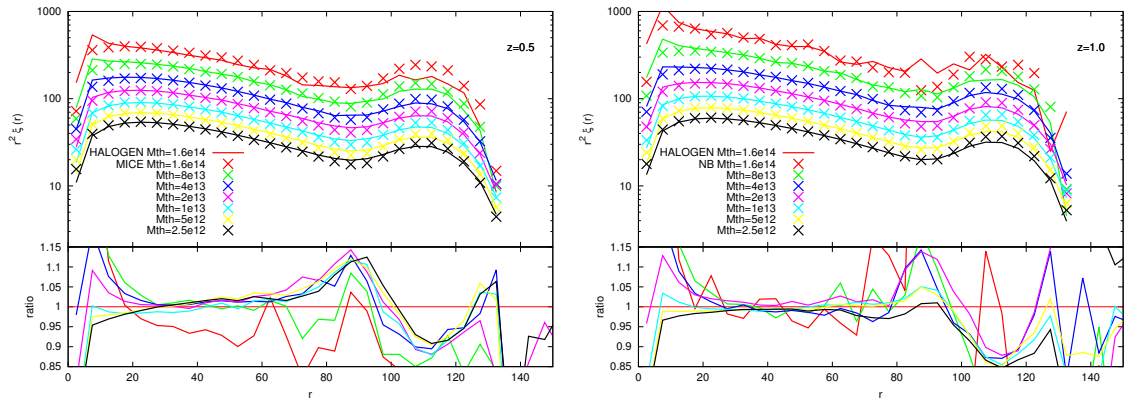

Figure 3.2: 2-point correlation function of MICE vs. HALOGEN halos in the simulation box at the snapshots $z = 0.5$ (left) and $z = 1.0$ (right). We show the different mass thresholds used during the fit.

shells ($\Delta z = 0.005 - 0.025$) of snapshots generated from a full N-Body simulation, and using the velocity of the particles to extrapolate their positions at the precise moment they cross the lightcone [164]. Remarkably, despite the great differences in the methodology, the angular correlation function from both lightcones shows very good agreement at all the interpolated redshifts.

3.2.2 Photometric Redshift

As already explained, a photometric survey can determine the redshift of a galaxy with a precision of $\sigma_z/(1+z) \approx 0.03$, depending on the sample and the estimating algorithm. Here, we use data with photo-z estimated by BPZ [165, 166]. The aim of this section is to apply this effect in the mock catalogues.

Data have been decided to be binned in 8 $z_{\rm ph}$-bins of width 0.05 between $z_{\rm ph} = 0.6$ and $z_{\rm ph} = 1.0$. The same algorithm that provides the $z_{\rm ph}$ estimates an error in that measurement and $P(z_{\rm rsd}|i)$ can be estimated. This is, the probability of a galaxy having redshift $z_{\rm rsd}$ given that it lies in the $z_{\rm ph}$-bin i. These probabilities are shown in Figure 3.4 (crosses) and can be fitted by a Gaussian with width $\sigma^i = s^i \cdot (1+z)$ (although not explicitly shown here), whose values are shown in Table 3.1.

A naive way to apply this to our catalogues would be to add a Gaussian random

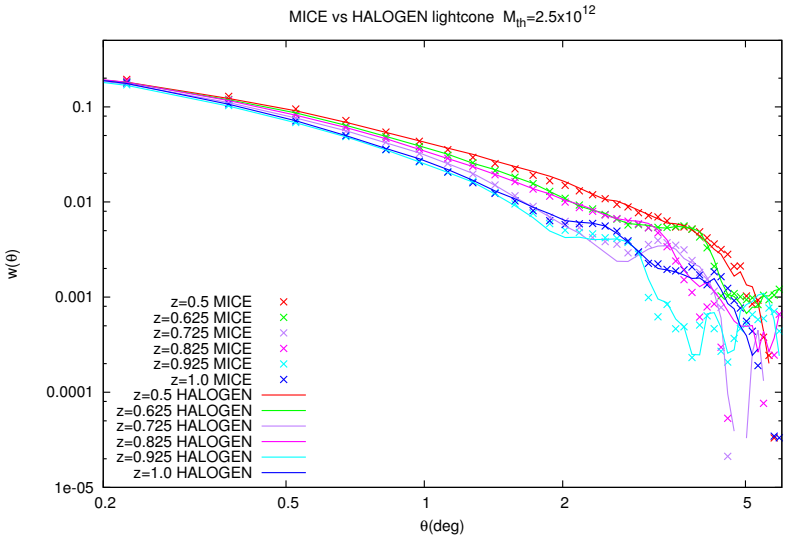

Figure 3.3: Angular correlation function of halos from the MICE (crosses) and HALOGEN (lines) lightcones. The different curves correspond to different redshift bins with width $\Delta z_{\rm rsd} = 0.5$ and centred at the indicated $z_{\rm rsd}$.

number with the same width σ^i to our $z_{\rm rsd}$, i.e.

$$z_{\rm ph} = z_{\rm rsd} + \Delta_{\rm ph}(z_{\rm rsd}) \cdot (1 + z_{\rm rsd}) \cdot R_{\rm gauss}(0, 1) \tag{3.3}$$

with

$$\Delta_{\rm ph}(z) = s_{\rm data}^{{\rm bin}(z)} \tag{3.4}$$

being $R_{\rm gauss}(0, 1)$ a Gaussian random number with mean 0 and standard deviation 1, and ${\rm bin}(z)$ a discrete function giving i between 1 and 8 according to Table 3.1. For $z_{\rm rsd} > 1.0$ and $z_{\rm rsd} < 0.6$ we use, respectively, the value of $\Delta_{\rm ph}$ from the last and first bin.

However, as shown in Figure 3.4 (*cp*, dashed lines), this is far from reproducing the data. The reason is that, as σ^i is not independent of redshift (although s^i is flat in a certain z range), a width in a $P(z_{\rm rsd}|z_{\rm ph})$ distribution is not equivalent to a width in $P(z_{\rm ph}|z_{\rm rsd})$. This concept may be better understood with an illustration: a galaxy

bin-i	1	2	3	4	5	6	7	8	–	–	Ξ^2
z-range	[0.6,0.65)	[0.65,0.7)	[0.7,0.75)	[0.75,0.8)	[0.8,0.85)	[0.85,0.9)	[0.9,0.95)	[0.95,1.0)	[1.0,1.1)	[1.1,1.2)	–
$s^i_{\rm data}$	0.030	0.030	0.029	0.029	0.030	0.035	0.041	0.049	–	–	–
$\Delta^i_{\rm cp}$	0.030	0.030	0.029	0.029	0.030	0.035	0.041	0.049	–	–	0.612
$s^i_{\rm cp}$	0.031	0.034	0.039	0.042	0.044	0.044	0.044	0.043	–	–	–
$\Delta^i_{\rm opt}$	0.031	0.029	0.029	0.029	0.029	0.029	0.029	0.030	0.040	0.050	–
$s^i_{\rm opt}$	0.030	0.030	0.031	0.032	0.035	0.039	0.040	0.039	–	–	0.098

Table 3.1: Photo-z. Redshift intervals for the 8 z-bins (rows 2 and 1) and, below, measured (from $P(z_{\rm rsd}|i)$) and applied (in Equation 3.3) widths. First (row 3), we find the measured width in the data $s^i_{\rm data}$, then we present the input (Δ) and output (s) of two models: *cp* for which we take $\Delta^i_{\rm cp} = s^i_{\rm data}$ and *opt* for which $\Delta^i_{\rm opt}$ are free and set to minimise Ξ^2. For $\Delta^i_{\rm opt}$ we allow two additional z-bins. Finally, in the last column we show the Ξ^2 as defined in Equation 3.5 obtained for the two models.

that has been assigned $z_{\rm ph} = 0.65$ will be more likely coming from $z_{\rm rsd} = 0.8$ than from $z_{\rm rsd} = 0.5$, since the error applied at higher redshifts is bigger. This effect skews and widens the distribution. This is also seen in the widths $s^i_{\rm cp}$ measured from the $P(z_{\rm rsd}|i)$ distribution of the catalogs after applying this method (Table 3.1).

In order to improve this, in a second method, we vary the values of $\Delta^i_{\rm ph}$ and minimise

$$\Xi^2 = \sum_{i=1}^{8} \frac{(s^i_{\rm method} - s^i_{\rm data})^2}{(s^i_{\rm data})^2} \tag{3.5}$$

where, $s^i_{\rm method}$ is the measured width in the $P(z_{\rm rsd}|i)$ distribution after having applied $\Delta^i_{\rm method}$ in Equation 3.3.

Further, we allow different values of $\Delta_{\rm ph}$ in two additional bins, $z \in [1.0, 1.1)$ and $z \in [1.1, 1.2)$, as we find it helps minimizing Ξ^2. The best fit values $\Delta^i_{\rm opt}$ and the outcome $s^i_{\rm opt}$ are shown in the last two rows of Table 3.1. Note that $\Delta_{\rm opt}$ remains nearly flat in all the target redshift range, and that it is the contamination from higher redshifts what makes $s^i_{\rm opt}$ change with redshift.

The $P(z_{\rm rsd}|i)$ for this method has been also plotted in Figure 3.4 (*opt*, solid lines), showing an improvement with respect to the previous method. We fix this photo-z scheme for the rest of the results presented below.

3.2.3 Galaxies with HOD and HAM

So far, all the clustering measurements shown throughout the thesis were obtained from halo catalogues at a given mass threshold. But observed clustering is typi-

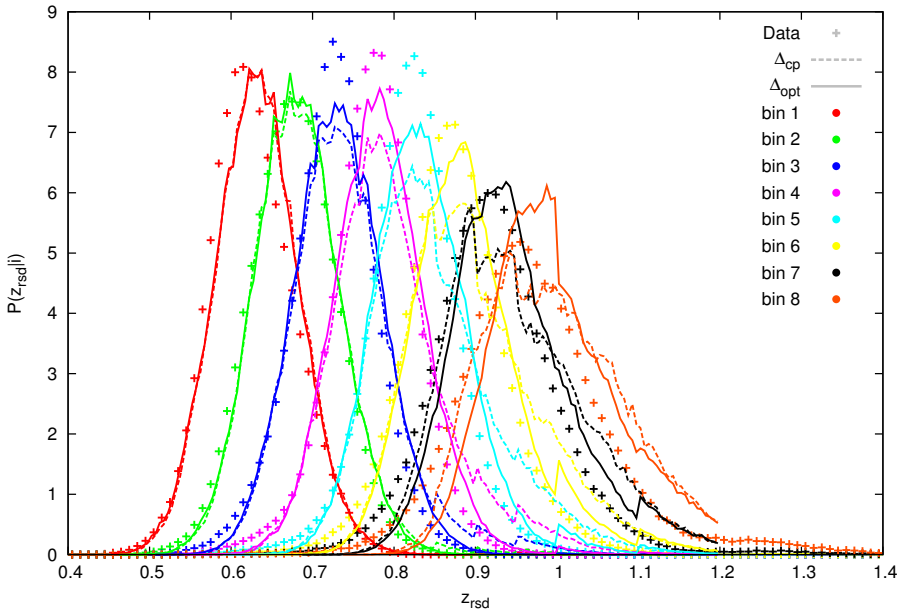

Figure 3.4: $z_{\rm rsd}$ probability distribution in each $z_{\rm ph}$-bin i for the data and the two photo-z schemes described in the text and applied to the mock catalogues.

cally measured from galaxy catalogues with a magnitude limited sample with the associated selection effects and, more generally, with redshift-dependent colour and magnitude cuts.

One could assume that the most massive halo in a simulation would correspond to the most luminous galaxy in the observations and that we could do a one-to-one mapping in rank order. This is certainly very optimistic and we need to add a scatter in the Luminosity-Mass relation ($L - M$) that will decrease the clustering for a magnitude-limit sample. This idea presented here is the basis of the Halo Abundance Matching (HAM) method [76–79].

HALOGEN was designed to only deal with main halos, neglecting subhalos (Section 2.3.2 & Section 2.1.2). This limits the potential of HAM, as we can not use its natural extension to subhalos SHAM, where there is more freedom in the physics modelling (see e.g. [167]).

Nevertheless, we already argued (Section 2.1.2) the possibility of adding substructure to a main halo catalogue with a Halo Occupation Distribution (HOD) scheme. We

know that halos can host more than one galaxy, especially massive halos which represent galaxy clusters and host tens of galaxies. If we attribute a number of galaxies $N_{\rm gal}$ which is an increasing function of the halo mass (M_h) to a halo mock catalogue, the clustering will be enhanced, since massive halos will be over-represented (as occurs in reality). This is the basis of the HOD methods [67–75].

We just presented two models that need to be implemented to a halo catalogue to get a realistic galaxy catalogue. The details of these methods need to be matched to observations via parameter fitting. This process can be particularly difficult if one aims at having a general model that serves for any sample with any magnitude and colour cut at any redshift (e.g. [73]). Additionally, the HOD implementation will determine the small scale clustering corresponding to the correlation between galaxies of the same halo. This is the 1-halo term according to the halo model [168], as opposed to the 2-halo term, which is relevant at large scales and corresponds to correlation between galaxies of different halos.

Here, we aim at presenting the first end-to-end galaxy mock catalogue set for the Y1-LSS selected sample. Hence, we start from the simplest model that can match the large scale clustering and number density. This will consist in applying a HAM with a one-parameter $L-M$ scatter when the halo-clustering is higher than the data and a one-parameter HOD when the halo-clustering is lower than the data.

As we are only interested in a particular sample, we use $M^{\rm gal}$ as a proxy for luminosity, and the HAM scatter is modelled as

$$\log M^{\rm gal} = \log M_{\rm h} + \gamma \cdot R_{\rm gauss}(0,1) \tag{3.6}$$

with γ a free parameter, indicated the dispersion in *dex* (decimal exponent units).

For the HOD, we always set a central galaxy at the centre of the halo and $N_{\rm sat}$ satellite galaxies following a NFW profile [66] where $N_{\rm sat}$ is a Poisson draw of the halo mass divided by the free parameter M_1:

$$N_{\rm sat}(M_h) = R_{\rm Poisson}\left(\frac{M_h}{M_1}\right) \tag{3.7}$$

We can find in the literature more complex $N_{\rm sat}(M_h)$ functions that include exponentials, exponentials by parts and error functions. However, we chose Equation 3.7 for simplicity in the fitting, finding valid results for the desired purpose. Moreover, studies using power-law HODs find best fit values for the exponent very close to unity [73], which leads to Equation 3.7.

The concentration relation needed for the NFW placement is determined by the mass following [169]. The velocities of the central galaxies are taken from the host halo, whereas the velocities of the satellite have an added dispersion following [170, 171]

$$\begin{aligned} v_{\rm sat} &= v_{\rm halo} + \sigma_v(M_h)\cdot R_{\rm gauss}(0,1) \\ \sigma_v(M_h) &= 476 f_{\rm vir}\left[\Delta_{\rm vir}E(z)^2\right]^{1/6}\left(\frac{M_h}{10^{15}M_\odot/h}\right)^{1/3} km/s \end{aligned} \tag{3.8}$$

where we use $f_{\rm vir}=0.9$ and $\Delta_{\rm vir}(z)=18\pi^2+82d(z)-39d(z)^2$ from spherical top-hat collapse theory, being $d(z)=1-\Omega(z)$ and $E(z)^2=H(z)^2/H_0^2$. Note that virial theorem together with Equation 1.9 already predicts $\sigma\propto M^{1/3}$.

All the galaxies contained in a halo with mass M_h are assigned with the same mass $M^{\rm gal}=M_h$.

This HOD-HAM process is done before constructing the lightcone and applying the photo-z, but measurements of the target $w^i(\theta)$ in the 8 $z_{\rm ph}$-bins and its associated χ^2 are performed after those processes:

$$\chi^2=\sum_{i=1}^{8}\sum_{0.1^\circ<\theta_j<1^\circ}\frac{(w^i_{\rm data}(\theta_j)-\bar{w}^i(\theta_j))^2}{\Delta w(\theta_j)^2} \tag{3.9}$$

Here, we use the same z-bins previously introduced in Table 3.1. The fit of this procedure is done from 8 catalogues as follows:

- In each $z_{\rm true}$-bin i apply either the HOD **or** the HAM scatter with one parameter (M_1^i or γ^i) depending on whether we need to enhance or reduce the clustering, respectively. The bin-1 value is also used for the low-z extension of the lightcone ($z_{\rm true}<0.6$) and the bin-8 for the upper part ($z_{\rm true}>1.0$).

- Apply the lightcone, mask and photo-z. This mixes galaxies drawn from different physics (γ^i/M_1^i), but this helps smoothing the transition between z-bins.
- Compute the mass threshold $M_{\rm th}^{\rm gal}(z_{\rm ph})$ (a proxy for a redshift-dependent magnitude cut) needed to match the $N(z)$ of data for each of the 8 catalogues. Compute the average of them and apply that threshold $\bar{M}_{\rm th}^{\rm gal}(z_{\rm ph})$ to all the catalogues.
- Compute the angular correlation function in the 8 $z_{\rm ph}$-bins for the 8 catalogues and measure their mean $\bar{w}(\theta)$ and standard deviation $\Delta w(\theta)$ to estimate χ^2 (Equation 3.9)
- If convergence of χ^2 is reached the fit is finished, otherwise, the processes is repeated with another set of parameters $\{M_1^i,\gamma^i\}$.

The resulting fitted parameters are shown in Table 3.2. These are used for the generation of the 8 HALOGEN-LAMPS catalogues whose statistics are shown in Figure 3.5 and Figure 3.6. Both number density and angular clustering show an excellent agreement with data. Moreover, we see that the galaxy mock catalogues (HALOGEN-LAMPS) represents a great improvement with respect to the halo catalogues (HALOGEN). Hence, the implementation of the HOD-HAM scheme appears necessary.

Finally, we remark that the dispersion γ found in the last three bins is large compared to the typical dispersions found in the literature [172, 173]. This is partially due to the sample selection and partially due to the modelling. Firstly, in those bins, the density field drops quickly (Figure 3.5), and low density (highly biased) samples typically present more dispersion. Moreover, the photo-z selection gets more contaminated (see the broadening in Figure 3.4, or $s^i_{\rm data}$ values), and what is meant to be a highly biased sample (especially due to the low density) may be selecting average galaxies from other redshift, needing more scatter to compensate. Finally, the halo catalogues have a mass resolution of $M = 2.5 \cdot 10^1 1 h^{-1} M_\odot$. Given the large dispersions that we are applying (over 2-3 orders of magnitudes), it is clear, we are lacking lower mass halos that would decrease the bias more efficiently. In fact, the bias barely changes after $\gamma \gtrsim 1.5$, clearly pointing towards the convenience of improving the mass resolution of the catalogue.

bin-i	1	2	3	4	5	6	7	8
$\log_{10}(M_1^i)$	13.4	13.6	14.2	14.5	14.0	–	–	–
γ^i	–	–	–	–	–	2.6	2.6	3.5

Table 3.2: HOD and HAM fitted parameters for the 8 z-bins (Table 3.1). M_1 is the mass scale of the HOD and γ the scatter in the $L-M$ relation in dex.

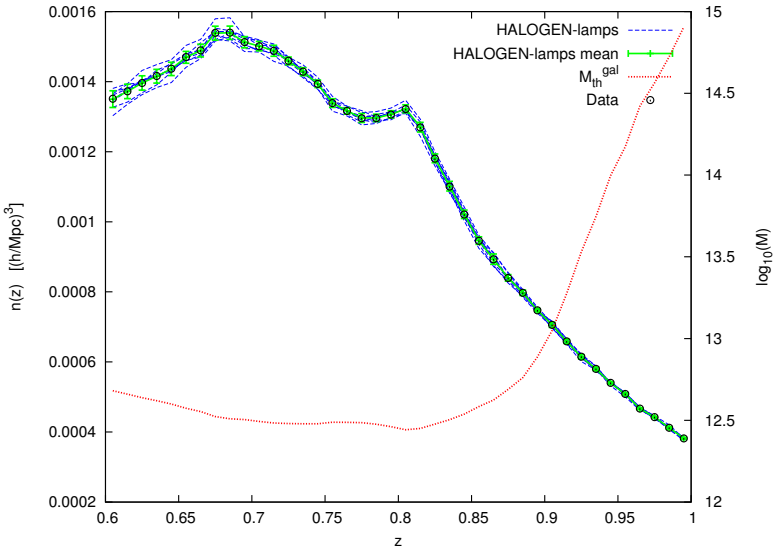

Figure 3.5: Left-axis: Number density of galaxies in units of $(h/Mpc)^3$ as a function of $z_{\rm ph}$ for data and galaxy mock catalogues. For the latter we show the mean and σ (HALOGEN-LAMPS mean), and individual curves (HALOGEN-LAMPS) over 8 realisations. Right-axis: Mass threshold $\bar{M}_{\rm th}^{\rm gal}$ used to get the HALOGEN-LAMPS catalogues, the value indicates the decimal logarithm of mass in $h^{-1}M_\odot$.

Certainly, as we improve our understanding on the data, we will improve the modelling and vice-versa. At the moment, in this section, we have constructed the first set of catalogues that reproduce the three main properties of the Y1-LSS sample, as shown in Figure 3.4, Figure 3.5 and Figure 3.6.

3.3 Results and Applications

Once we have a set of galaxy mock catalogues, we can use them for many applications:

- First, gain insight into the modelling and compare statistics with theoretical

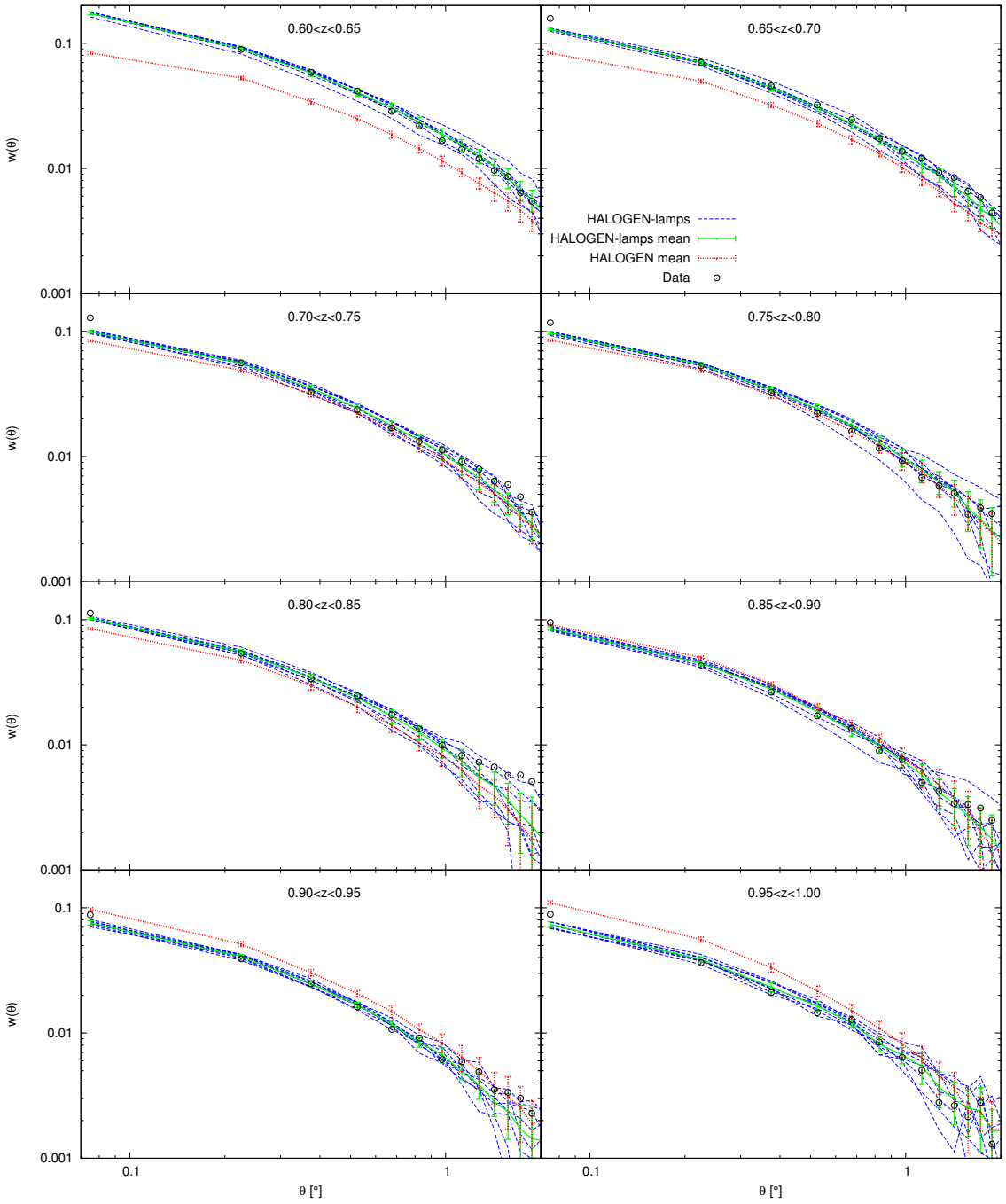

Figure 3.6: Angular correlation function in the 8 $z_{\rm ph}$-bins (as indicated in each panel). We compare the clustering of the Y1-LSS sample data with the halo mock catalogues (HALOGEN) and the galaxy mock catalogues (HALOGEN-LAMPS), showing the mean and standard deviation computed over 8 mock catalogues. Further, we show the clustering of each of the 8 galaxy mock catalogues that can be more directly comparable with the data. For all catalogues have been imposed the same rough $n(z)$

predictions (Section 3.3.1).

- Additionally, we can study the optimal methodology to extract the data (Section 3.3.2).
- Eventually, compute covariance matrices and set the uncertainty on the –BAO and other– measurements Section 3.3.3.

All the results presented in this section are provisional, and some of the figures presented here were based on a previous version of the catalogues (with a simpler photo-z modelling and only halos). But we want to emphasise the need and functionality of these catalogues in the process of the analysis and optimisation rather than presenting results, that will not be definitive until the optimisation in the sample is finished, the method fixed, and the results published by the collaboration.

3.3.1 Modelling Insight

Whereas in Section 3.2.3 we already compared the catalogues with data during the calibration, here we start by comparing them with purely theoretical models. This will help us understanding the models and their range of validity.

We show in Figure 3.7 a comparison of the clustering and its error with theoretical predictions done by the method explained in [174]. The theory part implemented the same bias $b(z)$, photo-z $P(z_{\rm rsd}|i)$ and number density $N(z)$ as the mock catalogues. We find a good agreement both for the mean and error on $w(\theta)$ for all $z_{\rm ph}$-bins, although, as expected, the theoretical predictions underestimate the errors. These errors represent the diagonal part of the covariance matrices. We leave for a future study the comparison of off-diagonal components.

Being able to model the data from simulations allows us to understand better the physics behind, and to control it in the simulation. For example, in the left panel of Figure 3.8 we study the exact effect of adding a photo-z to our catalogues in the clustering and the difference between the two photo-z models introduced in Section 3.2.2.

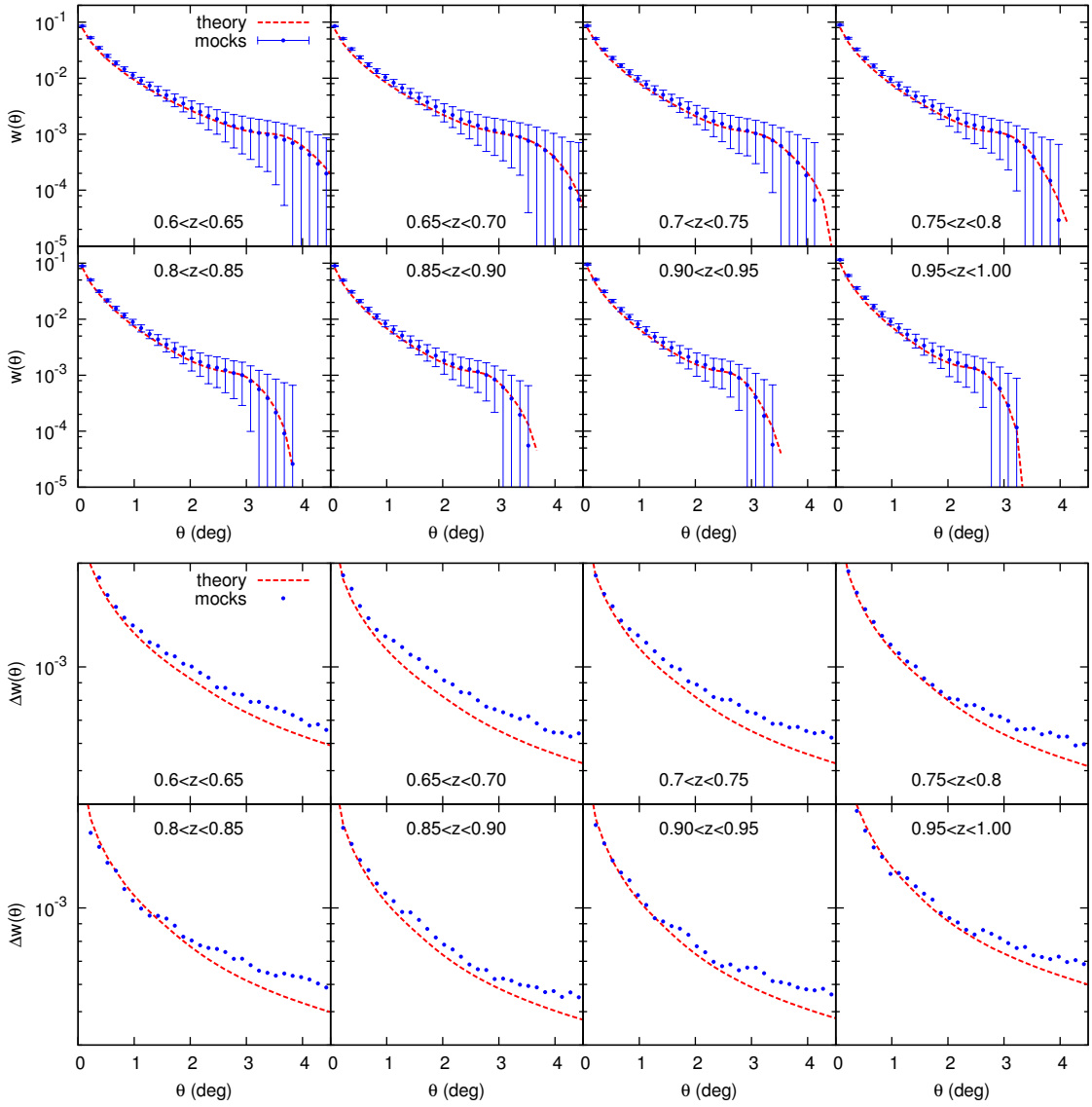

Figure 3.7: Comparison of estimation from theory and mock catalogues of the mean of the angular clustering (top) and the $1-\sigma$ error (bottom) for the 8 $z_{\rm ph}$-bins. 504 HALOGEN halo mock catalogues have been used for this estimations. Theory predictions provided by M. Crocce.

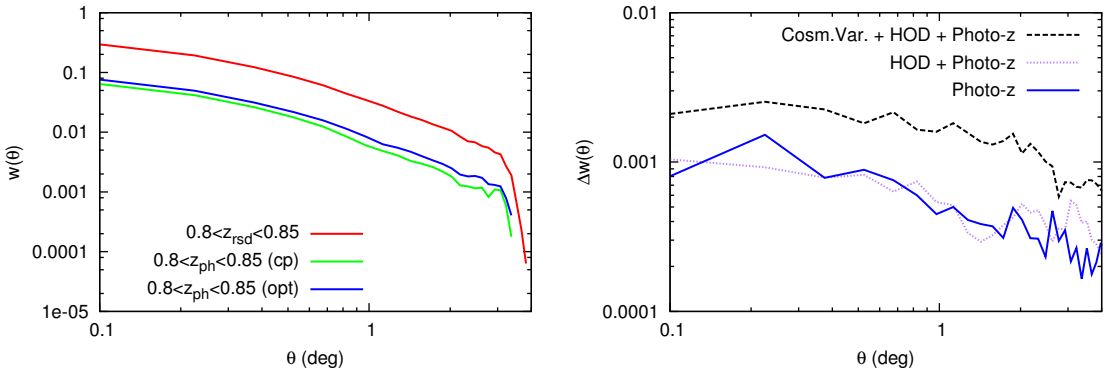

Figure 3.8: Studying the effects of photo-z. On the left we compare the clustering of a sample selected in $z_{\rm rsd}$ with a sample selected in $z_{\rm ph}$ following the two different methods explained in Section 3.2.2. All the catalogues were selected using the same number density. On the right we compare the error introduced in $w(\theta)$ by the photo-z, the HOD and cosmic variance (see text) for the $0.8 < z_{\rm ph} < 0.85$ bin.

As expected, the clustering decreases by adding the photo-z, because the density field is effectively smoothed along the line of sight and inhomogeneities appear less pronounced. The model labelled as *cp*, that presents a higher $\sigma_{\rm ph}$ (Table 3.1), reduces even more the clustering.

One of the motivations that we argued for the need of mock catalogues was to account for the interplay of systematics and cosmic variance. In the right panel of Figure 3.8, we show a comparison of the error induced in $w(\theta)$ by the photo-z, by the combination of photo-z and HOD, and the total error by the combination of photo-z, HOD and cosmic variance. This was computed from a) 8 different realisations of the photo-z on the same catalogue without HOD; b) 8 realisations of photo-z and HOD on the same catalogue, c) 8 different catalogues with the full implementation. The HOD here refers to the combination of HOD and HAM scatter as fixed by Table 3.2. Interestingly, we find that the error introduced by the photo-z seeds an important fraction ($\sim 0.3 - 0.5$) of the total error, whereas the HOD stochasticity introduces a negligible error.

3.3.2 Optimizing methodology

Another important application of mocks is the test of the methodology. The way we compress the data from a list of the coordinates of the full galaxy catalogue to a $\chi_{\rm BAO}$ measurement, will affect $\chi_{\rm BAO}$ itself and particularly $\Delta\chi_{\rm BAO}$. This can be analysed statistically with a large set of mock catalogues.

In the past sections we took 8 $z_{\rm ph}$-bins and an angular binning of 0.015° for the $w(\theta)$ as a starting point. The convenience of changing both bin sizes to optimise the precision in $\chi_{\rm BAO}$ is now under study. Preliminary results from theory suggest that a smaller θ-binning will increase the precision in the BAO and that $z_{\rm ph}$-bins can be widen without information loss while reducing the size of the covariance matrices. This needs to be confirmed or refuted by mock catalogues, since the validity of theoretical predictions at small θ-binning may be non-realistic.

In addition, a comparison of the different methods to extract the BAO information will be carried out in [175], where different proposed methods analyse both the data and the mock catalogues. This will include methods that analyse the clustering in 3D space ($\xi(s)$), the angular clustering in configuration space ($w(\theta)$) and in Fourier space (C_l).

Although most of the methods extract the BAO from the angular clustering because most of the information along the line of sight is lost, preliminary results of [176] show that combining carefully the 3D information one can recover $\chi_{\rm BAO}$ with similar precision as from the angular clustering with the advantage of reducing drastically the dataset. In [176] we study how the BAO information distributes with μ (cosine of the angle with respect to the line-of-sight), finding a non-negligible amount at $0.2 < \mu < 0.4$ that is neglected by angular clustering. The relative information in different μ intervals can be seen in Figure 3.9, where we see a well pronounced peak at $\mu < 0.2$ that fades at larger μ.

3.3.3 Uncertainty

Finally, the ultimate goal of the mock catalogues is to compute the covariance matrices of the correlation functions and the error bars of the BAO scale. Preliminary

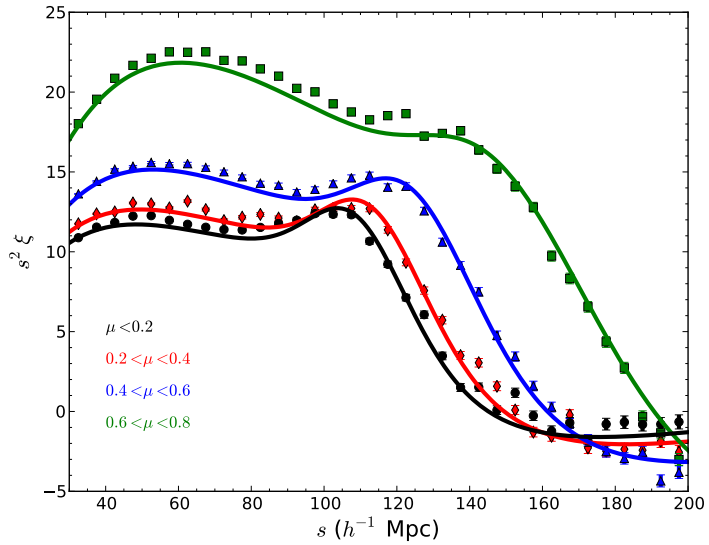

Figure 3.9: 3D correlation $\xi(s)$ integrated in μ intervals from [176]. Solid lines follow the theoretical models, whereas points are computed with 504 HALOGEN mock catalogues, error bars represent the standard deviation over $\sqrt{504}$.

results are shown in Figure 3.10, where we compare the correlation from the data (BPZ) with HALOGEN mock catalogues, a catalogue from an N-body simulation (Buzzard) and a theoretical model. The errors from the data and Buzzard are computed with a Jack-Knife algorithm. In the future we will see a similar figure with the errors estimated purely from the mock catalogues. The results are very encouraging and all the work done in the LSS working group is promising.

3.4 Conclusions

We have presented the first end-to-end set of galaxy mock catalogues for the Y1-LSS sample of the Dark Energy Survey. They have been designed to match the data in three statistics

- Photo-z distribution $P(z_{\rm rsd}|i)$ in the 8 $z_{\rm ph}$-bins.
- 1-point statistics: $n(z_{\rm ph})$

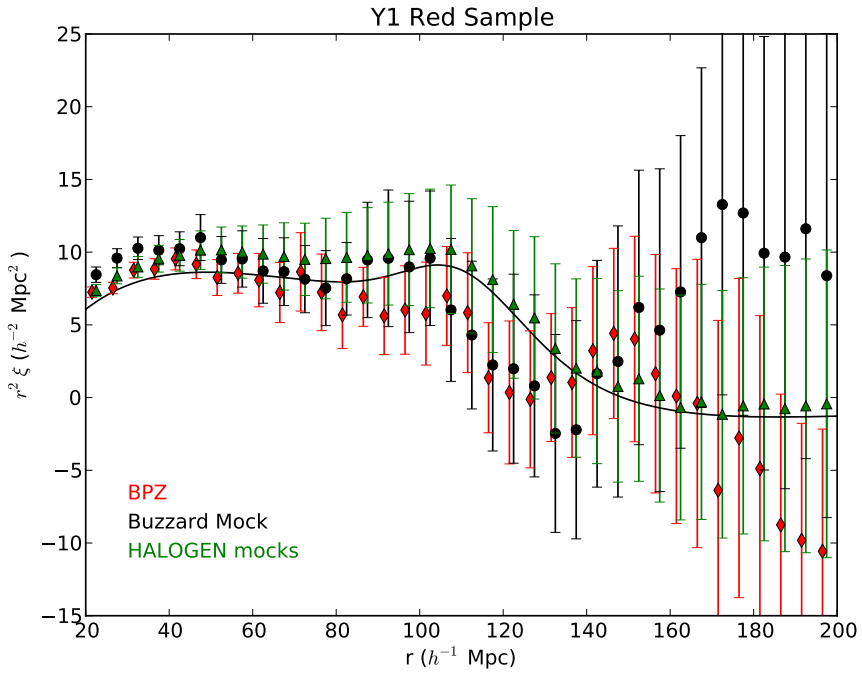

Figure 3.10: Preliminary 3D correlation function $\xi(r)$ from data (BPZ), HALOGEN catalogues, Buzzard catalogue (N-body simulation) and theoretical model (solid line). HALOGEN points represent the mean and standard deviation from 504 halo mock catalogues. Errors in both Buzzard and BPZ are estimated via Jack-Knife. Figure provided by A. J. Ross

- 2-point statistics: $w^i(\theta)$ in the 8 $z_{\rm ph}$-bins.

With respect to previous methodology explained in Chapter 2, HALOGEN-LAMPS presented here suppose the implementation of 3 new features:

- Lightcone. The parameters of HALOGEN are fitted with several snapshots (at $z = 0.0, 0.5, 1.0, 1.5$) and then interpolated at intermediate redshifts. The lightcone is constructed by superposition of z-shells of $\Delta z = 0.05$ at the range of interest.

- Photo-z. We implement a Gaussian model for the photo-z, whose applied widths $\Delta^i_{\rm ph}$ are optimised to match the observed widths σ^i.

- Galaxies. In order to adapt the clustering of halos to the observed galaxy clustering we introduced two implementations: a HOD and a HAM with a $L - M$ scatter. The HAM scatter mixes halos of different bias in the same magnitude-limited sample, reducing the clustering. The HOD places more galaxies in the most massive halos enhancing the clustering.

During the construction of this set, we acknowledged some improvements to be done in the future. First, the transition in z is abrupt when we change from one snapshot to another, and we are implementing a version where the lightcone is constructed in a unique snapshot where the growth factors depend on $r(z)$ (Equation 3.2). Secondly, we still find some differences in the photo-z implemented and the one measured (Figure 3.4). It appears that drawing galaxies from $z_{\rm rsd} < 1.2$ is not complete for the $8^{\rm th}$ bin, and the lightcone must be completed until $z_{\rm rsd} = 1.4$. Finally, a more realistic galaxy-halo distribution would be desired if we want to extend the use of the catalogues for other samples or other physics (e.g. cluster physics). For this, we need to implement together the two models –HOD and the scatter HAM– and a more complex HOD capable of matching the magnitude-dependent, colour-dependent, as well as the small scale clustering. A natural option for a future implementation is applying the method presented in [73] that was used for the reference MICE simulation. Regardless of whether we use this method or a new one, it appears that we will need to improve the mass resolution. In a future release we plan to have complete

catalogues up to $M_h \sim 3 \cdot 10^{11} M_\odot/h$. By having constructed a first complete set of galaxy mock catalogue, we learned many features that can be improved at once in the next release.

The analysis of the data is a complex process that needs several iterations in which the sample, methodology, and mock catalogues tell each other how they need to be modified until a degree of convergence is reached.

In the horizon, we would like to include more physics that can be measured with DES or other experiments, including weak-lensing, cluster physics, cross-correlation with CMB or intensity mapping, higher order statistics, etc. For all of this, the basics of the method might need to change, for example, we may need to resolve more the low-scale physics with a gravity solver beyond 2LPT.

Nonetheless, we have shown that this set of HALOGEN-LAMPS catalogues has been essential for the analysis of the Y1-LSS sample, contributing to the optimisation of the methodology, learning about the physics behind the data, and eventually will determine the error bars in the BAO measurements. There is still work ahead to finish the analysis of Y1-LSS data and everything remains at a preliminary stage, but Figure 3.10 is a promise for a *land ahoy!* shout.

Closure

In this thesis, I presented the research carried out during my PhD in the field of Large Scale Structure of the Universe. It connects on the one side simulations and on the other side observations. In Chapter 1, I studied the suitability of certain halo finders and merger tree builders in the field of N-Body simulations. During Chapter 2, I developed a technique to generate approximate halo mock catalogs: HALOGEN. Finally, in Chapter 3 that technique was applied to the analysis of observed data from a specific galaxy survey: the Dark Energy Survey. I extensively presented the conclusions of each chapter in Section 1.6, Section 2.7 and Section 3.4 respectively, but I synthesize here some of the most general conclusions, together with a wider outlook.

In Chapter 1, I showed that we need to carefully select/design the tools we use in the simulation pipeline accordingly to the application. Specifically, for accurate merger trees, we need halo finders able to trace halos when crossing the centre of another halo. Achieving this can be aided by tracking algorithms or phase-space finders. It is also desirable to have merger tree builders able to correct for halo finder flaws by, for example, skipping one snapshot.

In Chapter 2, I argued for the need for a new generation of tools for the massive production of synthetic galaxy catalogs. HALOGEN was shown to be a powerful tool able to produce halo catalogs with the correct 1 and 2-point statistics at large scales. It consists on a single-parameter bias routine applied to a 2LPT density field, with an analytic mass function and a velocity rescaling. I also presented a comparison of

practically all existing approximate methods (at that time) and discussed how each of them is best suited for different applications, depending on the trade-off between accuracy, simplicity, amount of resources and versatility.

All my previous work culminates in Chapter 3, where HALOGEN is extended and adapted to realistic observations: HALOGEN-LAMPS. It includes the realization of a lightcone, an optimized method to simulate photometric redshift and the inclusion of galaxies with a HOD-HAM scheme. HALOGEN-LAMPS catalogs are shown to match the observed Y1-LSS DES sample in angular clustering, number density and photo-z dispersion. Preliminary results showed how HALOGEN-LAMPS catalogs are being used to understand the modelling and improve the methodology in the data extraction. Finally, we showed tentative results in Figure 3.10, where we proved the potential of HALOGEN to set error bars on the clustering and, eventually, measure the error in $\chi_{\rm BAO}$ and its associated covariance matrices.

Although no final results can be claimed until the selection and analysis of the Y1-LSS sample concludes and the results are published, things are converging fast and we expect to have a new BAO detection soon. But this is only the first year data out of 5 year of DES observations. During that period a new point will appear in the $d_M(z)$ diagram presented in Figure 2, eventually settling at the unexplored BAO region of $z = 1$, perhaps solving some of the current puzzles in Cosmology, or maybe posing new questions.

We have shown that precision Cosmology with Large Scale Structure is possible, but not necessarily easy. The field of fast generation of mock catalogs is now boosting, and will need soon to deal simultaneously with increasing volumes, higher precision in measurements, more statistics to reproduce and larger covariance matrices to be estimated. There is a lot of work ahead to be performed by the scientific community, and a new generation of tools is needed.

Data analysis, theory and simulations must interact with each other, opening new windows to explore and new horizons to reach. The cosmological revolution goes on, supported by the work of countless cosmologist around the world doing their share, and little by little widening the knowledge as a collective mind.

Epílogo

En esta tesis he presentado el trabajo de investigación realizado durante mi doctorado en el campo de la Estructura a Gran Escala del Universo. En él se conecta las simulaciones por un lado, y las observaciones por el otro lado. En el Capítulo 1 estudiamos la idoneidad de varios *Halo Finders* y *Merger Tree builders* en el campo de las simulaciones de N-cuerpos, o simulaciones N-Body. A lo largo del Capítulo 2 desarrollé una nueva técnica para generar catálogos de halos de manera aproximada: HALOGEN. Por último, en el Capítulo 3, ese método es utilizado para el análisis de datos observacionales de un cartografiado específico: el Dark Energy Survey (DES). Las conclusiones de cada uno de los capítulos han sido ampliamente debatidas en Sección 1.6, Sección 2.7 y Sección 3.4. No obstante, resumiré brevemente las conclusiones principales a continuación.

En el Capítulo 1 esudiamos la necesidad seleccionar o diseñar de manera cuidadosa las herramientas que utilizamos en nuestras simulaciones, dependiendo de las aplicaciones para las que vayan a ser utilizadas. Más específicamente, para generar *Merger Trees* de manera rigurosa, necesitamos *Halo Finders* que puedan seguir el rastro de los halos incluso cuando están cruzando el centro de otro halo. Para conseguir esto, un buen método es usar *Halo Finders* que utilicen el espacio de fases o que cuenten con un algoritmo de seguimiento. Además, podemos mejorar la calidad de los *Merger Trees* utilizando *Merger Tree builders* que puedan corregir los errores del *Halo Finder*, por ejemplo, permitiendo la omisión de un *snapshot*.

En el Capítulo 2 expliqué la necesidad de una nueva generación de herramientas para la producción masiva de catálogos simulados de galaxias. A continuación, desarrollamos el método HALOGEN y mostramos su capacidad para producir catálogos de halos con las estadísticas de 1-punto y 2-puntos adecuadas (a escalas grandes). HALOGEN consiste en un mecanismo de *bias* cosmológico con un único parámetro libre, aplicado a un campo generado con 2LPT (Teoría Lagrangiana de Perturbaciones a 2º orden), la creación de masas de halos mediante funciones de masas analíticas y un reajuste de la velocidad. También presenté una comparación de, prácticamente, todos los métodos aproximados existentes y debatí cómo cada uno de ellos se adecúa a diferentes aplicaciones, dependiendo del balance entre precisión, simplicidad, recursos y versatilidad.

Todo el trabajo anterior culmina en el Capítulo 3, donde extiendo las funciones de HALOGEN y el método es adaptado para observaciones realistas de galaxias: HALOGEN-LAMPS. En este capítulo, añadimos la construcción de un cono de luz, un método optimizado para simular el redshift fotométrico, y la inclusión de galaxias con un método HOD-HAM. Demostramos que los catálogos de HALOGEN-LAMPS tienen las mismas características que la muestra observacional de DES Y1-LSS en la correlación angular, densidad de número de galaxias y distribución de photo-z, todo ello como función del redshift z. Con unos resultados preliminares presentados en la Sección 3.3, mostré cómo los catálogos de HALOGEN-LAMPS están siendo utilizados para entender mejor la modelización de los datos y mejorar la metodología en la extracción de información con los datos. Por último, en la Figura 3.10 mostramos resultados provisionales donde probamos el potencial de HALOGEN-LAMPS para establecer barras de error en la función de correlación y, en el futuro, medir el error en $\chi_{\rm BAO}$ y sus matrices de covarianza asociadas.

Aunque los resultados finales no estarán listos hasta que la selección y análisis de la muestra Y1-LSS haya acabado, y los resultados hayan sido publicados por la colaboración; la situación converge rápidamente y esperamos obtener una nueva detección del BAO pronto. Pero esto es sólo el análisis de los datos del primer año, de

los 5 años programados en DES. Durante ese tiempo, un nuevo punto aparecerá en el diagrama $d_M(z)$ que introjimos en la Figura 2. Con nuevos datos y el correspondiente análisis, este punto se irá asentando en torno al área inexplorada de $z \approx 1$. Quizás ese nuevo punto nos ayude a resolver algunos de los enigmas actuales de la Cosmolgía o, quizás, plantee nuevas preguntas.

A lo largo de esta tesis he mostrado que la Cosmología de Precisión con la Estructura a Gran Escala es posible, pero no necesariamente fácil. El campo de la generación rápida de catálogos simulados de galaxias está reciviendo un fuerte estímulo. Pronto requerirá lidiar simultáneamente con volúmenes más grandes, más precisión en las medidas, ser capaz de reproducir más estadísticas y estimar matrices de covarianza más grandes. Aún queda en el camino mucho trabajo, al que la comunidad deberá enfrentarse para construir una nueva generación de herramientas de simulaciones cosmológicas.

El análisis de datos, las simulaciones y la teoría deberán interactuar los uno con los otros, abriendo nuevas ventanas para explorar nuevos horizontes. La Revolución Cosmológica continúa, contando con el apoyo de innumerables cosmólogos a lo largo y ancho del planeta, contribuyendo con su grano de arena, y poco a poco ensanchando el conocimiento como una mente colectiva.

Bibliography

[1] S. Avila, A. Knebe, F. R. Pearce, A. Schneider, C. Srisawat, P. A. Thomas, P. Behroozi, P. J. Elahi, J. Han, Y.-Y. Mao, J. Onions, V. Rodriguez-Gomez, and D. Tweed. SUSSING MERGER TREES: the influence of the halo finder. MNRAS, 441:3488–3501, July 2014.

[2] S. Avila, S. G. Murray, A. Knebe, C. Power, A. S. G. Robotham, and J. Garcia-Bellido. HALOGEN: a tool for fast generation of mock halo catalogues. MNRAS, 450:1856–1867, June 2015.

[3] C.-H. Chuang, C. Zhao, F. Prada, E. Munari, S. Avila, A. Izard, F.-S. Kitaura, M. Manera, P. Monaco, S. Murray, A. Knebe, C. G. Scóccola, G. Yepes, J. Garcia-Bellido, F. A. Marín, V. Müller, R. Skibba, M. Crocce, P. Fosalba, S. Gottlöber, A. A. Klypin, C. Power, C. Tao, and V. Turchaninov. nIFTy cosmology: Galaxy/halo mock catalogue comparison project on clustering statistics. MNRAS, 452:686–700, September 2015.

[4] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, and et al. Planck 2015 results. XIII. Cosmological parameters. *ArXiv e-prints*, February 2015.

[5] G. Bertone, D. Hooper, and J. Silk. Particle dark matter: evidence, candidates and constraints. Phys. Rep., 405:279–390, January 2005.

[6] L. Amendola and S. Tsujikawa. *Dark Energy: Theory and Observations*. Cambridge University Press, 2010.

[7] A. Linde. Inflationary Cosmology after Planck 2013. *ArXiv e-prints*, February 2014.

[8] J. Lee and E. Komatsu. Bullet Cluster: A Challenge to ΛCDM Cosmology. ApJ, 718:60–65, July 2010.

[9] S. Chongchitnan and J. Silk. Primordial non-Gaussianity and extreme-value statistics of galaxy clusters. Phys. Rev. D, 85(6):063508, March 2012.

[10] A. Klypin, A. V. Kravtsov, O. Valenzuela, and F. Prada. Where Are the Missing Galactic Satellites? ApJ, 522:82–92, September 1999.

[11] Planck Collaboration, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, N. Bartolo, and et al. Planck 2015 results. XI. CMB power spectra, likelihoods, and robustness of parameters. *ArXiv e-prints*, July 2015.

[12] J. S. Bullock, A. V. Kravtsov, and D. H. Weinberg. Reionization and the Abundance of Galactic Satellites. ApJ, 539:517–521, August 2000.

[13] A. Drlica-Wagner, K. Bechtol, E. S. Rykoff, E. Luque, A. Queiroz, Y.-Y. Mao, R. H. Wechsler, J. D. Simon, B. Santiago, B. Yanny, E. Balbinot, S. Dodelson, A. Fausti Neto, D. J. James, T. S. Li, M. A. G. Maia, J. L. Marshall, A. Pieres, K. Stringer, A. R. Walker, T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. Brooks, E. Buckley-Geer, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, M. Crocce, L. N. da Costa, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, T. F. Eifler, A. E. Evrard, D. A. Finley, B. Flaugher, P. Fosalba, J. Frieman, E. Gaztanaga, D. W. Gerdes, D. Gruen, R. A. Gruendl, G. Gutierrez, K. Honscheid, K. Kuehn, N. Kuropatkin, O. Lahav, P. Martini, R. Miquel, B. Nord, R. Ogando, A. A. Plazas, K. Reil, A. Roodman, M. Sako, E. Sanchez, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Tucker, V. Vikram, W. Wester, Y. Zhang, J. Zuntz, and DES Collaboration. Eight Ultra-faint Galaxy Candidates Discovered in Year Two of the Dark Energy Survey. ApJ, 813:109, November 2015.

[14] D. Kraljic and S. Sarkar. How rare is the Bullet Cluster (in a ΛCDM universe)? *Journal of Cosmology and Astroparticle Physics*, 4:050, April 2015.

[15] I. Harrison and P. Coles. Testing cosmology with extreme galaxy clusters. MNRAS, 421:L19–L23, March 2012.

[16] A. Rassat, J.-L. Starck, P. Paykari, F. Sureau, and J. Bobin. Planck CMB anomalies: astrophysical and cosmological secondary effects and the curse of masking. *Journal of Cosmology and Astroparticle Physics*, 8:006, August 2014.

[17] R. Amanullah, C. Lidman, D. Rubin, G. Aldering, P. Astier, K. Barbary, M. S. Burns, A. Conley, K. S. Dawson, S. E. Deustua, M. Doi, S. Fabbro, L. Faccioli, H. K. Fakhouri, G. Folatelli, A. S. Fruchter, H. Furusawa, G. Garavini, G. Goldhaber, A. Goobar, D. E. Groom, I. Hook, D. A. Howell, N. Kashikawa, A. G. Kim, R. A. Knop, M. Kowalski, E. Linder, J. Meyers, T. Morokuma, S. Nobili, J. Nordin, P. E. Nugent, L. Östman, R. Pain, N. Panagia, S. Perlmutter, J. Raux, P. Ruiz-Lapuente, A. L. Spadafora, M. Strovink, N. Suzuki, L. Wang, W. M. Wood-Vasey, N. Yasuda, and T. Supernova Cosmology Project. Spectra and Hubble Space Telescope Light Curves of Six Type Ia Supernovae at 0.511 ¡ z ¡ 1.12 and the Union2 Compilation. ApJ, 716:712–738, June 2010.

[18] Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. XVI. Cosmological parameters. A&A, 571:A16, November 2014.

[19] J. Frieman and Dark Energy Survey Collaboration. The Dark Energy Survey: Overview. In *American Astronomical Society Meeting Abstracts 221*, volume 221 of *American Astronomical Society Meeting Abstracts*, page 335.01, January 2013.

[20] R. Laureijs, J. Amiaux, S. Arduini, J. . Augueres, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, and et al. Euclid Definition Study Report. *ArXiv e-prints 1110.3193*, October 2011.

[21] P. J. E. Peebles. *Principles of Physical Cosmology*. Princeton Theories in Physics. Princeton University Press, 1993.

[22] D. W. Hogg. Distance measures in cosmology. *ArXiv Astrophysics e-prints*, May 1999.

[23] A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, B. Leibundgut, M. M. Phillips, D. Reiss, B. P. Schmidt, R. A. Schommer, R. C. Smith, J. Spyromilio, C. Stubbs, N. B. Suntzeff, and J. Tonry. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. AJ, 116:1009–1038, September 1998.

[24] S. Perlmutter, G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, A. Goobar, D. E. Groom, I. M. Hook, A. G. Kim, M. Y. Kim, J. C. Lee, N. J. Nunes, R. Pain, C. R. Pennypacker, R. Quimby, C. Lidman, R. S. Ellis, M. Irwin, R. G. McMahon, P. Ruiz-Lapuente, N. Walton, B. Schaefer, B. J. Boyle, A. V. Filippenko, T. Matheson, A. S. Fruchter, N. Panagia, H. J. M. Newberg, W. J. Couch, and T. S. C. Project. Measurements of Ω and Λ from 42 High-Redshift Supernovae. ApJ, 517:565–586, June 1999.

[25] P. J. E. Peebles and J. T. Yu. Primeval Adiabatic Perturbation in an Expanding Universe. ApJ, 162:815, December 1970.

[26] R. A. Sunyaev and Y. B. Zeldovich. Small-Scale Fluctuations of Relic Radiation. Ap&SS, 7:3–19, April 1970.

[27] S. Cole, W. J. Percival, J. A. Peacock, P. Norberg, C. M. Baugh, C. S. Frenk, I. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, M. Colless, C. Collins, W. Couch, N. J. G. Cross, G. Dalton, V. R. Eke, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, K. Glazebrook, C. Jackson, A. Jenkins, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, B. A. Peterson, W. Sutherland, and K. Taylor. The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications. MNRAS, 362:505–534, September 2005.

[28] D. J. Eisenstein, I. Zehavi, D. W. Hogg, R. Scoccimarro, M. R. Blanton, R. C. Nichol, R. Scranton, H.-J. Seo, M. Tegmark, Z. Zheng, S. F. Anderson, J. Annis, N. Bahcall, J. Brinkmann, S. Burles, F. J. Castander, A. Connolly, I. Csabai, M. Doi, M. Fukugita, J. A. Frieman, K. Glazebrook, J. E. Gunn, J. S. Hendry, G. Hennessy, Z. Ivezić, S. Kent, G. R. Knapp, H. Lin, Y.-S. Loh, R. H. Lupton, B. Margon, T. A. McKay, A. Meiksin, J. A. Munn, A. Pope, M. W. Richmond, D. Schlegel, D. P. Schneider, K. Shimasaku, C. Stoughton, M. A. Strauss, M. SubbaRao, A. S. Szalay, I. Szapudi, D. L. Tucker, B. Yanny, and D. G. York. Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. ApJ, 633:560–574, November 2005.

[29] F. Beutler, C. Blake, M. Colless, D. H. Jones, L. Staveley-Smith, L. Campbell, Q. Parker, W. Saunders, and F. Watson. The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant. MNRAS, 416:3017–3032, October 2011.

[30] C. Blake, E. A. Kazin, F. Beutler, T. M. Davis, D. Parkinson, S. Brough, M . Colless, C. Contreras, W. Couch, S. Croom, D. Croton, M. J. Drinkwater, K. Forster, D. Gilbank, M. Gladders, K. Glazebrook, B. Jelliffe, R. J. Jurek, I.-H. Li, B. Madore, D. C. Martin, K. Pimbblet, G. B. Poole, M. Pracy, R. Sharp, E. Wisnioski, D. Woods, T. K. Wyder, and H. K. C. Yee. The WiggleZ Dark Energy Survey: mapping the distance-redshift relation with baryon acoustic oscillations. MNRAS, 418:1707–1724, December 2011.

[31] L. Anderson, E. Aubourg, S. Bailey, D. Bizyaev, M. Blanton, A. S. Bolton, J. Brinkmann, J. R. Brownstein, A. Burden, A. J. Cuesta, L. A. N. da Costa, K. S. Dawson, R. de Putter, D. J. Eisenstein, J. E. Gunn, H. Guo, J.-C. Hamilton, P. Harding, S. Ho, K. Honscheid, E. Kazin, D. Kirkby, J.-P. Kneib, A. Labatie, C. Loomis, R. H. Lupton, E. Malanushenko, V. Malanushenko, R. Mandelbaum, M. Manera, C. Maraston, C. K. McBride, K. T. Mehta, O. Mena, F. Montesano, D. Muna, R. C. Nichol, S. E. Nuza, M. D. Olmstead, D. Oravetz, N. Padmanabhan, N. Palanque-Delabrouille, K. Pan, J. Parejko, I. Pâris, W. J. Percival, P. Petitjean, F. Prada, B. Reid, N. A. Roe, A. J.

Ross, N. P. Ross, L. Samushia, A. G. Sánchez, D. J. Schlegel, D. P. Schneider, C. G. Scóccola, H.-J. Seo, E. S. Sheldon, A. Simmons, R. A. Skibba, M. A. Strauss, M. E. C. Swanson, D. Thomas, J. L. Tinker, R. Tojeiro, M. V. Magaña, L. Verde, C. Wagner, D. A. Wake, B. A. Weaver, D. H. Weinberg, M. White, X. Xu, C. Yèche, I. Zehavi, and G.-B. Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample. MNRAS, 427:3435–3467, December 2012.

[32] N. G. Busca, T. Delubac, J. Rich, S. Bailey, A. Font-Ribera, D. Kirkby, J.-M. Le Goff, M. M. Pieri, A. Slosar, É. Aubourg, J. E. Bautista, D. Bizyaev, M. Blomqvist, A. S. Bolton, J. Bovy, H. Brewington, A. Borde, J. Brinkmann, B. Carithers, R. A. C. Croft, K. S. Dawson, G. Ebelke, D. J. Eisenstein, J.-C. Hamilton, S. Ho, D. W. Hogg, K. Honscheid, K.-G. Lee, B. Lundgren, E. Malanushenko, V. Malanushenko, D. Margala, C. Maraston, K. Mehta, J. Miralda-Escudé, A. D. Myers, R. C. Nichol, P. Noterdaeme, M. D. Olmstead, D. Oravetz, N. Palanque-Delabrouille, K. Pan, I. Pâris, W. J. Percival, P. Petitjean, N. A. Roe, E. Rollinde, N. P. Ross, G. Rossi, D. J. Schlegel, D. P. Schneider, A. Shelden, E. S. Sheldon, A. Simmons, S. Snedden, J. L. Tinker, M. Viel, B. A. Weaver, D. H. Weinberg, M. White, C. Yèche, and D. G. York. Baryon acoustic oscillations in the Lyα forest of BOSS quasars. A&A, 552:A96, April 2013.

[33] A. Slosar, V. Iršič, D. Kirkby, S. Bailey, N. G. Busca, T. Delubac, J. Rich, É. Aubourg, J. E. Bautista, V. Bhardwaj, M. Blomqvist, A. S. Bolton, J. Bovy, J. Brownstein, B. Carithers, R. A. C. Croft, K. S. Dawson, A. Font-Ribera, J.-M. Le Goff, S. Ho, K. Honscheid, K.-G. Lee, D. Margala, P. McDonald, B. Medolin, J. Miralda-Escudé, A. D. Myers, R. C. Nichol, P. Noterdaeme, N. Palanque-Delabrouille, I. Pâris, P. Petitjean, M. M. Pieri, Y. Piškur, N. A. Roe, N. P. Ross, G. Rossi, D. J. Schlegel, D. P. Schneider, N. Suzuki, E. S. Sheldon, U. Seljak, M. Viel, D. H. Weinberg, and C. Yèche. Measurement of baryon acoustic oscillations in the Lyman-α forest fluctuations in BOSS data release 9. *Journal of Cosmology and Astroparticle Physics*, 4:026, April 2013.

[34] A. A. Penzias and R. W. Wilson. A Measurement of Excess Antenna Temperature at 4080 Mc/s. ApJ, 142:419–421, July 1965.

[35] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, and et al. Planck 2015 results. XV. Gravitational lensing. *ArXiv e-prints*, February 2015.

[36] R. K. Sachs and A. M. Wolfe. Perturbations of a Cosmological Model and Angular Variations of the Microwave Background. ApJ, 147:73, January 1967.

[37] G. F. Smoot, C. L. Bennett, A. Kogut, E. L. Wright, J. Aymon, N. W. Boggess, E. S. Cheng, G. de Amici, S. Gulkis, M. G. Hauser, G. Hinshaw, P. D. Jackson, M. Janssen, E. Kaita, T. Kelsall, P. Keegstra, C. Lineweaver, K. Loewenstein, P. Lubin, J. Mather, S. S. Meyer, S. H. Moseley, T. Murdock, L. Rokke, R. F. Silverberg, L. Tenorio, R. Weiss, and D. T. Wilkinson. Structure in the COBE differential microwave radiometer first-year maps. ApJ, 396:L1–L5, September 1992.

[38] G. Hinshaw, M. R. Nolta, C. L. Bennett, R. Bean, O. Doré, M. R. Greason, M. Halpern, R. S. Hill, N. Jarosik, A. Kogut, E. Komatsu, M. Limon, N. Odegard, S. S. Meyer, L. Page, H. V. Peiris, D. N. Spergel, G. S. Tucker, L. Verde, J. L. Weiland, E. Wollack, and E. L. Wright. Three-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Temperature Analysis. ApJS, 170:288–334, June 2007.

[39] W. H. Press and P. Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. ApJ, 187:425–438, February 1974.

[40] Dark Energy Survey Collaboration. The Dark Energy Survey Science Program.

[41] B. Sartoris, A. Biviano, C. Fedeli, J. G. Bartlett, S. Borgani, M. Costanzi, C. Giocoli, L. Moscardini, J. Weller, B. Ascaso, S. Bardelli, S. Maurogordato, and P. T. P. Viana. Next Generation Cosmology: Constraints from the Euclid Galaxy Cluster Survey. MNRAS, March 2016.

[42] M. Kowalski, D. Rubin, G. Aldering, R. J. Agostinho, A. Amadon, R. Amanullah, C. Balland, K. Barbary, G. Blanc, P. J. Challis, A. Conley, N. V. Connolly, R. Covarrubias, K. S. Dawson, S. E. Deustua, R. Ellis, S. Fabbro, V. Fadeyev, X. Fan, B. Farris, G. Folatelli, B. L. Frye, G. Garavini, E. L. Gates, L. Germany, G. Goldhaber, B. Goldman, A. Goobar, D. E. Groom, J. Haissinski, D. Hardin, I. Hook, S. Kent, A. G. Kim, R. A. Knop, C. Lidman, E. V. Linder, J. Mendez, J. Meyers, G. J. Miller, M. Moniez, A. M. Mourão, H. Newberg, S. Nobili, P. E. Nugent, R. Pain, O. Perdereau, S. Perlmutter, M. M. Phillips, V. Prasad, R. Quimby, N. Regnault, J. Rich, E. P. Rubenstein, P. Ruiz-Lapuente, F. D. Santos, B. E. Schaefer, R. A. Schommer, R. C. Smith, A. M. Soderberg, A. L. Spadafora, L.-G. Strolger, M. Strovink, N. B. Suntzeff, N. Suzuki, R. C. Thomas, N. A. Walton, L. Wang, W. M. Wood-Vasey, and J. L. Yun. Improved Cosmological Constraints from New, Old, and Combined Supernova Data Sets. ApJ, 686:749–778, October 2008.

[43] O. Lahav and A. R Liddle. The Cosmological Parameters 2014. *ArXiv e-prints*, January 2014.

[44] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. Nature, 435:629–636, June 2005.

[45] J.W Eastwood R.W Hockney. *Computer simulation using particles*. A. Hilger, special student ed edition, 1988.

[46] A. Klypin. Numerical Simulations in Cosmology I: Methods. *ArXiv Astrophysics e-prints*, May 2000.

[47] M. Kuhlen, M. Vogelsberger, and R. Angulo. Numerical simulations of the dark universe: State of the art and the next decade. *Physics of the Dark Universe*, 1:50–93, November 2012.

[48] Antony Lewis, Anthony Challinor, and Anthony Lasenby. Efficient computa-

tion of CMB anisotropies in closed FRW models. *Astrophys. J.*, 538:473–476, 2000.

[49] F. Moutarde, J.-M. Alimi, F. R. Bouchet, R. Pellat, and A. Ramani. Precollapse scale invariance in gravitational instability. ApJ, 382:377–381, December 1991.

[50] F. R. Bouchet, S. Colombi, E. Hivon, and R. Juszkiewicz. Perturbative Lagrangian approach to gravitational instability. A&A, 296:575, April 1995.

[51] Y. B. Zel'dovich. Gravitational instability: An approximate theory for large density perturbations. A&A, 5:84–89, March 1970.

[52] M. Crocce, S. Pueblas, and R. Scoccimarro. Transients from initial conditions in cosmological simulations. MNRAS, 373:369–381, November 2006.

[53] J. Barnes and P. Hut. A hierarchical O(N log N) force-calculation algorithm. Nature, 324:446–449, December 1986.

[54] A. W. Appel. An Efficient Program for Many-Body Simulation. *SIAM Journal on Scientific and Statistical Computing, vol. 6, no. 1, January 1985, p. 85-103.*, 6:85–103, January 1985.

[55] L. Hernquist. Performance characteristics of tree codes. ApJS, 64:715–734, August 1987.

[56] A. Klypin and J. Holtzman. Particle-Mesh code for cosmological simulations. *ArXiv Astrophysics e-prints*, December 1997.

[57] G. Efstathiou, M. Davis, S. D. M. White, and C. S. Frenk. Numerical techniques for large cosmological N-body simulations. ApJS, 57:241–260, February 1985.

[58] H. M. P. Couchman. Mesh-refined P3M - A fast adaptive N-body algorithm. ApJ, 368:L23–L26, February 1991.

[59] J. Dubinski, J. Kim, C. Park, and R. Humble. GOTPM: a parallel hybrid particle-mesh treecode. NewA, 9:111–126, February 2004.

[60] V. Springel. The cosmological simulation code GADGET-2. MNRAS, 364:1105–1134, December 2005.

[61] A. V. Kravtsov, A. A. Klypin, and A. M. Khokhlov. Adaptive Refinement Tree: A New High-Resolution N-Body Code for Cosmological Simulations. ApJS, 111:73–94, July 1997.

[62] R. Teyssier. Cosmological hydrodynamics with adaptive mesh refinement. A new high resolution code called RAMSES. A&A, 385:337–364, April 2002.

[63] A. Knebe, F. R. Pearce, H. Lux, Y. Ascasibar, P. Behroozi, J. Casado, C. C. Moran, J. Diemand, and et al. Structure finding in cosmological simulations: the state of affairs. MNRAS, 435:1618–1658, October 2013.

[64] C. Srisawat, A. Knebe, F. R. Pearce, A. Schneider, P. A. Thomas, P. Behroozi, K. Dolag, P. J. Elahi, J. Han, J. Helly, Y. Jing, I. Jung, J. Lee, Y.-Y. Mao, J. Onions, V. Rodriguez-Gomez, D. Tweed, and S. K. Yi. Sussing Merger Trees: The Merger Trees Comparison Project. MNRAS, 436:150–162, November 2013.

[65] A. Knebe, F. R. Pearce, P. A. Thomas, A. Benson, J. Blaizot, R. Bower, J. Carretero, F. J. Castander, A. Cattaneo, S. A. Cora, D. J. Croton, W. Cui, D. Cunnama, G. De Lucia, J. E. Devriendt, P. J. Elahi, A. Font, F. Fontanot, J. Garcia-Bellido, I. D. Gargiulo, V. Gonzalez-Perez, J. Helly, B. Henriques, M. Hirschmann, J. Lee, G. A. Mamon, P. Monaco, J. Onions, N. D. Padilla, C. Power, A. Pujol, R. A. Skibba, R. S. Somerville, C. Srisawat, C. A. Vega-Martínez, and S. K. Yi. nIFTy cosmology: comparison of galaxy formation models. MNRAS, 451:4029–4059, August 2015.

[66] J. F. Navarro, C. S. Frenk, and S. D. M. White. The Structure of Cold Dark Matter Halos. ApJ, 462:563, May 1996.

[67] Y. P. Jing, H. J. Mo, and G. Börner. Spatial Correlation Function and Pairwise Velocity Dispersion of Galaxies: Cold Dark Matter Models versus the Las Campanas Survey. ApJ, 494:1–12, February 1998.

[68] J. A. Peacock and R. E. Smith. Halo occupation numbers and galaxy bias. MNRAS, 318:1144–1156, November 2000.

[69] A. A. Berlind and D. H. Weinberg. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. ApJ, 575:587–616, August 2002.

[70] Z. Zheng, A. A. Berlind, D. H. Weinberg, A. J. Benson, C. M. Baugh, S. Cole, R. Davé, C. S. Frenk, N. Katz, and C. G. Lacey. Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies. ApJ, 633:791–809, November 2005.

[71] H. Guo, I. Zehavi, and Z. Zheng. A New Method to Correct for Fiber Collisions in Galaxy Two-point Statistics. ApJ, 756:127, September 2012.

[72] I. Zehavi, Z. Zheng, D. H. Weinberg, M. R. Blanton, N. A. Bahcall, A. A. Berlind, J. Brinkmann, J. A. Frieman, J. E. Gunn, R. H. Lupton, R. C. Nichol, W. J. Percival, D. P. Schneider, R. A. Skibba, M. A. Strauss, M. Tegmark, and D. G. York. Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. ApJ, 736:59, July 2011.

[73] J. Carretero, F. J. Castander, E. Gaztañaga, M. Crocce, and P. Fosalba. An algorithm to build mock galaxy catalogues using MICE simulations. MNRAS, 447:646–670, February 2015.

[74] S. A. Rodríguez-Torres, C.-H. Chuang, F. Prada, H. Guo, A. Klypin, P. Behroozi, C. H. Hahn, J. Comparat, G. Yepes, A. D. Montero-Dorta, J. R. Brownstein, C. Maraston, C. K. McBride, J. Tinker, S. Gottlöber, G. Favole, Y. Shu, F.-S. Kitaura, A. Bolton, R. Scoccimarro, L. Samushia, D. Schlegel, D. P. Schneider, and D. Thomas. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Modeling the clustering and halo occupation distribution of BOSS-CMASS galaxies in the Final Data Release. *ArXiv e-prints*, September 2015.

[75] R. A. Skibba and R. K. Sheth. A halo model of galaxy colours and clustering in the Sloan Digital Sky Survey. MNRAS, 392:1080–1091, January 2009.

[76] C. Conroy, R. H. Wechsler, and A. V. Kravtsov. Modeling Luminosity-

dependent Galaxy Clustering through Cosmic Time. ApJ, 647:201–214, August 2006.

[77] P. S. Behroozi, C. Conroy, and R. H. Wechsler. A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for 0 ¡ z ¡ 4. ApJ, 717:379–403, July 2010.

[78] S. Trujillo-Gomez, A. Klypin, J. Primack, and A. J. Romanowsky. Galaxies in ΛCDM with Halo Abundance Matching: Luminosity-Velocity Relation, Baryonic Mass-Velocity Relation, Velocity Function, and Clustering. ApJ, 742:16, November 2011.

[79] H. Guo, Z. Zheng, P. S. Behroozi, I. Zehavi, C.-H. Chuang, J. Comparat, G. Favole, S. Gottloeber, A. Klypin, F. Prada, S. A. Rodríguez-Torres, D. H. Weinberg, and G. Yepes. Modelling Galaxy Clustering: Halo Occupation Distribution versus Subhalo Matching. MNRAS, April 2016.

[80] C. M. Baugh. A primer on hierarchical galaxy formation: the semi-analytical approach. *Reports on Progress in Physics*, 69:3101–3156, 2006.

[81] A. Knebe, S. R. Knollmann, S. I. Muldrew, F. R. Pearce, and et al. Haloes gone MAD: The Halo-Finder Comparison Project. MNRAS, 415:2293–2318, August 2011.

[82] D. J. Croton, V. Springel, S. D. M. White, G. De Lucia, C. S. Frenk, L. Gao, A. Jenkins, G. Kauffmann, J. F. Navarro, and N. Yoshida. The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. MNRAS, 365:11–28, January 2006.

[83] R. S. Somerville, P. F. Hopkins, T. J. Cox, B. E. Robertson, and L. Hernquist. A semi-analytic model for the co-evolution of galaxies, black holes and active galactic nuclei. MNRAS, 391:481–506, December 2008.

[84] P. Monaco, F. Fontanot, and G. Taffoni. The MORGANA model for the rise of galaxies and active nuclei. MNRAS, 375:1189–1219, March 2007.

[85] B. M. B. Henriques, P. A. Thomas, S. Oliver, and I. Roseboom. Monte Carlo Markov Chain parameter estimation in semi-analytic models of galaxy formation. MNRAS, 396:535–547, June 2009.

[86] A. J. Benson, S. Borgani, G. De Lucia, M. Boylan-Kolchin, and P. Monaco. Convergence of galaxy properties with merger tree temporal resolution. MNRAS, 419:3590–3603, February 2012.

[87] W. H. Press and P. Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. ApJ, 187:425–438, February 1974.

[88] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser. Excursion set mass functions for hierarchical Gaussian fluctuations. ApJ, 379:440–460, October 1991.

[89] F. Jiang and F. C. van den Bosch. Generating Merger Trees for Dark Matter Haloes: A Comparison of Methods. *ArXiv e-prints*, November 2013.

[90] B. F. Roukema, P. J. Quinn, and B. A. Peterson. Spectral Evolution of Merging/Accreting Galaxies. In G. L. Chincarini, A. Iovino, T. Maccacaro, and D. Maccagni, editors, *Observational Cosmology*, volume 51 of *Astronomical Society of the Pacific Conference Series*, page 51, January 1993.

[91] C. Lacey and S. Cole. Merger rates in hierarchical models of galaxy formation. MNRAS, 262:627–649, June 1993.

[92] A. Knebe, N. I. Libeskind, F. Pearce, P. Behroozi, J. Casado, K. Dolag, R. Dominguez-Tenreiro, P. Elahi, H. Lux, S. I. Muldrew, and J. Onions. Galaxies going MAD: the Galaxy-Finder Comparison Project. MNRAS, 428:2039–2052, January 2013.

[93] J. Onions, A. Knebe, F. R. Pearce, S. I. Muldrew, H. Lux, S. R. Knollmann, Y. Ascasibar, P. Behroozi, P. Elahi, J. Han, M. Maciejewski, M. E. Merchán, M. Neyrinck, A. N. Ruiz, M. A. Sgró, V. Springel, and D. Tweed. Subhaloes going Notts: the subhalo-finder comparison project. MNRAS, 423:1200–1214, June 2012.

[94] J. Onions, Y. Ascasibar, P. Behroozi, J. Casado, P. Elahi, J. Han, A. Knebe, H. Lux, M. E. Merchán, S. I. Muldrew, M. Neyrinck, L. Old, F. R. Pearce, D. Potter, A. N. Ruiz, M. A. Sgró, D. Tweed, and T. Yue. Subhaloes gone Notts: spin across subhaloes and finders. MNRAS, 429:2739–2747, March 2013.

[95] P. J. Elahi, J. Han, H. Lux, Y. Ascasibar, P. Behroozi, A. Knebe, S. I. Muldrew, J. Onions, and F. Pearce. Streams going Notts: the tidal debris finder comparison project. *submitted to MNRAS, arXiv:1305.2448*, June 2013.

[96] A. Knebe, F. R. Pearce, H. Lux, Y. Ascasibar, and et al. Structure Finding in Cosmological Simulations: The State of Affairs. *submitted to MNRAS, ArXiv:1304.0585*, April 2013.

[97] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. The evolution of large-scale structure in a universe dominated by cold dark matter. ApJ, 292:371–394, May 1985.

[98] V. Springel. The cosmological simulation code GADGET-2. MNRAS, 364:1105–1134, December 2005.

[99] E. Komatsu and et al. Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation. 192:18, February 2011.

[100] S. P. D. Gill, A. Knebe, and B. K. Gibson. The evolution of substructure - I. A new identification method. MNRAS, 351:399–409, June 2004.

[101] S. R. Knollmann and A. Knebe. AHF: Amiga's Halo Finder. ApJS, 182:608–624, June 2009.

[102] J. Han, Y. P. Jing, H. Wang, and W. Wang. Resolving subhaloes' lives with the Hierarchical Bound-Tracing algorithm. MNRAS, 427:2437–2449, December 2012.

[103] P. S. Behroozi, R. H. Wechsler, and H.-Y. Wu. The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. ApJ, 762:109, January 2013.

[104] V. Springel, S. D. M. White, G. Tormen, and G. Kauffmann. Populating a cluster of galaxies - I. Results at [formmu2]z=0. MNRAS, 328:726–750, December 2001.

[105] P. S. Behroozi, R. H. Wechsler, H.-Y. Wu, M. T. Busha, A. A. Klypin, and J. R. Primack. Gravitationally Consistent Halo Catalogs and Merger Trees for Precision Cosmology. ApJ, 763:18, January 2013.

[106] S. I. Muldrew, F. R. Pearce, and C. Power. The accuracy of subhalo detection. MNRAS, 410:2617–2624, February 2011.

[107] Y. Wang, F. R. Pearce, A. Knebe, A. Schneider, C. Srisawat, D. Tweed, I. Jung, J. Han, J. Helly, J. Onions, P. J. Elahi, P. A. Thomas, P. Behroozi, S. K. Yi, V. Rodriguez-Gomez, Y.-Y. Mao, Y. Jing, and W. Lin. Sussing Merger Trees: Stability and Convergence. *ArXiv e-prints*, April 2016.

[108] P. S. Behroozi, R. H. Wechsler, Y. Lu, O. Hahn, M. T. Busha, A. Klypin, and J. R. Primack. Mergers and Mass Accretion for Infalling Halos Both End Well Outside Cluster Virial Radii. *ArXiv e-prints*, October 2013.

[109] F. J. Castander, O. Ballester, A. Bauer, L. Cardiel-Sas, J. Carretero, R. Casas, J. Castilla, M. Crocce, M. Delfino, M. Eriksen, and et al. The PAU camera and the PAU survey at the William Herschel Telescope. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 8446 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 6, September 2012.

[110] K. S. Dawson, D. J. Schlegel, C. P. Ahn, S. F. Anderson, É. Aubourg, S. Bailey, R. H. Barkhouser, J. E. Bautista, A. Beifiori, and et al. The Baryon Oscillation Spectroscopic Survey of SDSS-III. AJ, 145:10, January 2013.

[111] M. Levi, C. Bebek, T. Beers, R. Blum, R. Cahn, D. Eisenstein, B. Flaugher, K. Honscheid, R. Kron, O. Lahav, P. McDonald, N. Roe, D. Schlegel, and representing the DESI collaboration. The DESI Experiment, a whitepaper for Snowmass 2013. *ArXiv e-prints 1308.0847*, August 2013.

[112] LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. J. Asztalos, T. S. Axelrod, and et al. LSST Science Book, Version 2.0. *ArXiv e-prints*, December 2009.

[113] M. Manera, R. Scoccimarro, and W. J. et al. Percival. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: a large sample of mock galaxy catalogues. MNRAS, 428:1036–1054, January 2013.

[114] P. Coles and B. Jones. A lognormal model for the cosmological mass distribution. MNRAS, 248:1–13, January 1991.

[115] R. Scoccimarro and R. K. Sheth. PTHALOS: a fast method for generating mock galaxy distributions. MNRAS, 329:629–640, January 2002.

[116] P. Monaco, T. Theuns, and G. Taffoni. The pinocchio algorithm: pinpointing orbit-crossing collapsed hierarchical objects in a linear density field. MNRAS, 331:587–608, April 2002.

[117] P. Monaco, E. Sefusatti, S. Borgani, M. Crocce, P. Fosalba, R. K. Sheth, and T. Theuns. An accurate tool for the fast generation of dark matter halo catalogues. MNRAS, 433:2389–2402, August 2013.

[118] F.-S. Kitaura, G. Yepes, and F. Prada. Modelling baryon acoustic oscillations with perturbation theory and stochastic halo biasing. MNRAS, 439:L21–L25, March 2014.

[119] S. Tassev, M. Zaldarriaga, and D. J. Eisenstein. Solving large scale structure in ten easy steps with COLA. *Journal of Cosmology and Astroparticle Physics*, 6:36, June 2013.

[120] M. White, J. L. Tinker, and C. K. McBride. Mock galaxy catalogues using the quick particle mesh method. MNRAS, 437:2594–2606, January 2014.

[121] C.-H. Chuang, F.-S. Kitaura, F. Prada, C. Zhao, and G. Yepes. EZmocks: extending the Zel'dovich approximation to generate mock galaxy catalogues with accurate clustering statistics. MNRAS, 446:2621–2628, January 2015.

[122] Y. Feng, M.-Y. Chu, and U. Seljak. FastPM: a new scheme for fast simulations of dark matter and halos. *ArXiv e-prints*, March 2016.

[123] S. R. Knollmann and A. Knebe. AHF: Amiga's Halo Finder. ApJS, 182:608–624, June 2009.

[124] A. Klypin, G. Yepes, S. Gottlober, F. Prada, and S. Hess. MultiDark simulations: the story of dark matter halo concentrations and density profiles. *ArXiv e-prints 1411.4001*, November 2014.

[125] Planck Collaboration. Planck 2013 results. XVI. Cosmological parameters. A&A, 571:A16, November 2014.

[126] P. Fosalba, M. Crocce, E. Gaztañaga, and F. J. Castander. The MICE grand challenge lightcone simulation - I. Dark matter clustering. MNRAS, 448:2987–3000, April 2015.

[127] M. Crocce, F. J. Castander, E. Gaztañaga, P. Fosalba, and J. Carretero. The MICE Grand Challenge lightcone simulation - II. Halo and galaxy catalogues. MNRAS, 453:1513–1530, October 2015.

[128] P. Fosalba, E. Gaztañaga, F. J. Castander, and M. Crocce. The MICE Grand Challenge light-cone simulation - III. Galaxy lensing mocks from all-sky lensing maps. MNRAS, 447:1319–1332, February 2015.

[129] R. Scoccimarro. Transients from initial conditions: a perturbative analysis. MNRAS, 299:1097–1118, October 1998.

[130] M. C. Neyrinck. Quantifying distortions of the Lagrangian dark-matter mesh in cosmology. MNRAS, 428:141–153, January 2013.

[131] V. Sahni and S. Shandarin. Accuracy of Lagrangian approximations in voids. MNRAS, 282:641–645, September 1996.

[132] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser. Excursion set mass functions for hierarchical Gaussian fluctuations. ApJ, 379:440–460, October 1991.

[133] S. G. Murray, C. Power, and A. S. G. Robotham. HMFcalc: An online tool for calculating dark matter halo mass functions. *Astronomy and Computing*, 3:23–34, November 2013.

[134] W. A. Watson, I. T. Iliev, A. D'Aloisio, A. Knebe, P. R. Shapiro, and G. Yepes. The halo mass function through the cosmic ages. MNRAS, 433:1230–1245, August 2013.

[135] D. Alonso. CUTE solutions for two-point correlation functions from large cosmological datasets. *ArXiv e-prints 1210.1833*, October 2012.

[136] J. L. Tinker, D. H. Weinberg, Z. Zheng, and I. Zehavi. On the Mass-to-Light Ratio of Large-Scale Structure. ApJ, 631:41–58, September 2005.

[137] R. W. Hockney and J. W. Eastwood. *Computer simulation using particles*. 1988.

[138] P. Colín, A. A. Klypin, and A. V. Kravtsov. Velocity Bias in a Λ Cold Dark Matter Model. ApJ, 539:561–569, August 2000.

[139] P. J. E. Peebles. Statistics of the distribution of galaxies. In J. Ehlers, J. J. Perry, and M. Walker, editors, *Ninth Texas Symposium on Relativistic Astrophysics*, volume 336 of *Annals of the New York Academy of Sciences*, pages 161–171, February 1980.

[140] J. N. Fry. Cosmological density fluctuations and large-scale structure From N-point correlation functions to the probability distribution. ApJ, 289:10–17, February 1985.

[141] W. C. Saslaw. *The Distribution of the Galaxies*. February 2000.

[142] Y. P. Jing. Correcting for the Alias Effect When Measuring the Power Spectrum Using a Fast Fourier Transform. ApJ, 620:559–563, February 2005.

[143] E. Hubble. The Distribution of Extra-Galactic Nebulae. ApJ, 79:8, January 1934.

[144] V. Wild, J. A. Peacock, O. Lahav, E. Conway, S. Maddox, I. K. Baldry, C. M. Baugh, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, M. Colless, C. Collins, W. Couch, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, I. Lewis, S. Lumsden, D. Madgwick, P. Norberg, B. A. Peterson, W. Sutherland, and K. Taylor. The 2dF Galaxy Redshift Survey: stochastic relative biasing between galaxy populations. MNRAS, 356:247–269, January 2005.

[145] F.-S. Kitaura and S. Heß. Cosmological structure formation with augmented Lagrangian perturbation theory. MNRAS, 435:L78–L82, August 2013.

[146] C. Zhao, F.-S. Kitaura, C.-H. Chuang, F. Prada, G. Yepes, and C. Tao. Halo mass distribution reconstruction across the cosmic web. MNRAS, 451:4266–4276, August 2015.

[147] H. Gil-Marín, J. Noreña, L. Verde, W. J. Percival, C. Wagner, M. Manera, and D. P. Schneider. The power spectrum and bispectrum of SDSS DR11 BOSS galaxies - I. Bias and gravity. MNRAS, 451:539–580, July 2015.

[148] M. Bolzonella, J.-M. Miralles, and R. Pelló. Photometric redshifts based on standard SED fitting procedures. A&A, 363:476–492, November 2000.

[149] N. Benítez. Bayesian Photometric Redshift Estimation. ApJ, 536:571–583, June 2000.

[150] A. E. Firth, O. Lahav, and R. S. Somerville. Estimating photometric redshifts with artificial neural networks. MNRAS, 339:1195–1202, March 2003.

[151] Dark Energy Survey Collaboration, T. Abbott, F. B. Abdalla, S. Allam, J. Aleksić, A. Amara, D. Bacon, E. Balbinot, M. Banerji, K. Bechtol, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, J. Blazek, S. Dodelson, C. Bonnett, D. Brooks, S. Bridle, R. J. Brunner, E. Buckley-Geer, D. L. Burke, D. Capozzi, G. B. Caminha, J. Carlsen, A. Carnero-Rosell, M. Carollo, M. Carrasco-Kind, J. Carretero, F. J. Castander, L. Clerkin, T. Collett, C. Conselice, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, T. M.

Davis, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. Drlica-Wagner, J. Etherington, J. Estrada, A. E. Evrard, J. Fabbri, D. A. Finley, B. Flaugher, P. Fosalba, R. J. Foley, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, D. A. Goldstein, D. Gruen, R. A. Gruendl, P. Guarnieri, G. Gutierrez, W. Hartley, K. Honscheid, B. Jain, D. J. James, T. Jeltema, S. Jouvel, R. Kessler, A. King, D. Kirk, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, M. Lima, H. Lin, M. A. G. Maia, M. Makler, M. Manera, C. Maraston, J. L. Marshall, P. Martini, R. G. McMahon, P. Melchior, A. Merson, C. J. Miller, R. Miquel, J. J. Mohr, X. Morice-Atkinson, K. Naidoo, E. Neilsen, R. C. Nichol, B. Nord, R. Ogando, F. Ostrovski, A. Palmese, A. Papadopoulos, H. Peiris, J. Peoples, A. A. Plazas, W. J. Percival, S. L. Reed, A. K. Romer, A. Roodman, A. Ross, E. Rozo, E. S. Rykoff, I. Sadeh, M. Sako, C. Sánchez, E. Sanchez, B. Santiago, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, E. Sheldon, M. Smith, R. C. Smith, M. Soares-Santos, F. Sobreira, M. Soumagnac, E. Suchyta, M. Sullivan, M. Swanson, G. Tarle, J. Thaler, D. Thomas, R. C. Thomas, D. Tucker, J. D. Vieira, V. Vikram, A. R. Walker, R. H. Wechsler, W. Wester, J. Weller, L. Whiteway, H. Wilcox, B. Yanny, Y. Zhang, and J. Zuntz. The Dark Energy Survey: more than dark energy - an overview. MNRAS, March 2016.

[152] K. Batygin and M. E. Brown. Evidence for a Distant Giant Planet in the Solar System. AJ, 151:22, February 2016.

[153] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Physical Review Letters*, 116(6):061102, February 2016.

[154] M. Soares-Santos, R. Kessler, E. Berger, J. Annis, D. Brout, E. Buckley-Geer, H. Chen, P. S. Cowperthwaite, H. T. Diehl, Z. Doctor, A. Drlica-Wagner, B. Farr, D. A. Finley, B. Flaugher, R. J. Foley, J. Frieman, R. A. Gruendl, K. Herner, D. Holz, H. Lin, J. Marriner, E. Neilsen, A. Rest, M. Sako, D. Scolnic, F. Sobreira, A. R. Walker, W. Wester, B. Yanny, T. M. C. Abbott, F. B. Abdalla, F. B. Abdalla, S. Allam, R. Armstrong, M. Banerji, A. Benoit-

Lévy, R. A. Bernstein, E. Bertin, D. A. Brown, D. L. Burke, D. Capozzi, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, S. B. Cenko, R. Chornock, M. Crocce, C. B. D'Andrea, C. B. D'Andrea, L. N. da Costa, S. Desai, J. P. Dietrich, M. R. Drout, T. F. Eifler, J. Estrada, A. E. Evrard, S. Fairhurst, E. Fernandez, J. Fischer, W. Fong, P. Fosalba, D. B. Fox, C. L. Fryer, J. Garcia-Bellido, E. Gaztanaga, D. W. Gerdes, D. A. Goldstein, D. Gruen, G. Gutierrez, K. Honscheid, D. J. James, I. Karliner, D. Kasen, S. Kent, N. Kuropatkin, K. Kuehn, O. Lahav, T. S. Li, M. Lima, M. A. G. Maia, R. Margutti, P. Martini, T. Matheson, R. G. McMahon, B. D. Metzger, C. J. Miller, R. Miquel, J. J. Mohr, R. C. Nichol, B. Nord, R. Ogando, J. Peoples, A. A. Plazas, E. Quataert, A. K. Romer, A. Roodman, A. Roodman, E. S. Rykoff, E. S. Rykoff, E. Sanchez, V. Scarpine, R. Schindler, M. Schubnell, I. Sevilla-Noarbe, E. Sheldon, M. Smith, N. Smith, R. C. Smith, A. Stebbins, P. J. Sutton, M. E. C. Swanson, G. Tarle, J. Thaler, R. C. Thomas, D. L. Tucker, V. Vikram, R. H. Wechsler, and J. Weller. A Dark Energy Camera Search for an Optical Counterpart to the First Advanced LIGO Gravitational Wave Event GW150914. *ArXiv e-prints*, February 2016.

[155] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Localization and broadband follow-up of the gravitational-wave transient GW150914. *ArXiv e-prints*, February 2016.

[156] M. E. Brown and K. Batygin. Observational constraints on the orbit and location of Planet Nine in the outer solar system. *ArXiv e-prints*, March 2016.

[157] J. Kwan, C. Sanchez, J. Clampitt, J. Blazek, M. Crocce, B. Jain, J. Zuntz, A. Amara, M. Becker, G. Bernstein, C. Bonnett, J. DeRose, S. Dodelson, T. Eifler, E. Gaztanaga, T. Giannantonio, D. Gruen, W. Hartley, T. Kacprzak, D. Kirk, E. Krause, N. MacCrann, R. Miquel, Y. Park, A. Ross, E. Rozo, E. Rykoff, E. Sheldon, M. A. Troxel, R. Wechsler, T. Abbott, F. Abdalla, S. Allam, A. Benoit-Lévy, D. Brooks, D. Burke, A. Carnero Rosell, M. Carrasco Kind, C. Cunha, C. D'Andrea, L. da Costa, S. Desai, H. T. Diehl, J. Dietrich, P. Doel, A. Evrard, E. Fernandez, D. Finley, B. Flaugher, P. Fosalba, J. Frie-

man, D. Gerdes, R. Gruendl, G. Gutierrez, K. Honscheid, D. James, M. Jarvis, K. Kuehn, O. Lahav, M. Lima, M. Maia, J. Marshall, P. Martini, P. Melchior, J. Mohr, R. Nichol, B. Nord, A. Plazas, K. Reil, K. Romer, A. Roodman, E. Sanchez, V. Scarpine, I. Sevilla, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. Swanson, G. Tarle, D. Thomas, V. Vikram, and A. Walker. Cosmology from large scale galaxy clustering and galaxy-galaxy lensing with Dark Energy Survey Science Verification data. *ArXiv e-prints*, April 2016.

[158] T. Kacprzak, D. Kirk, O. Friedrich, A. Amara, A. Refregier, L. Marian, J. P. Dietrich, E. Suchyta, J. Aleksić, D. Bacon, M. R. Becker, C. Bonnett, S. L. Bridle, C. Chang, T. F. Eifler, W. Hartley, E. M. Huff, E. Krause, N. MacCrann, P. Melchior, A. Nicola, S. Samuroff, E. Sheldon, M. A. Troxel, J. Weller, J. Zuntz, T. M. C. Abbott, F. B. Abdalla, R. Armstrong, A. Benoit-Lévy, R. A. Bernstein, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, M. Crocce, C. B. D'Andrea, L. N. da Costa, S. Desai, H. T. Diehl, A. E. Evrard, A. Fausti Neto, B. Flaugher, P. Fosalba, J. Frieman, D. W. Gerdes, D. A. Goldstein, D. Gruen, R. A. Gruendl, G. Gutierrez, K. Honscheid, D. J. James, K. Kuehn, N. Kuropatkin, O. Lahav, M. Lima, M. March, J. L. Marshall, P. Martini, C. J. Miller, R. Miquel, J. J. Mohr, R. C. Nichol, B. Nord, A. A. Plazas, A. K. Romer, A. Roodman, E. S. Rykoff, E. Sanchez, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, M. E. C. Swanson, G. Tarle, D. Thomas, V. Vikram, A. R. Walker, and Y. Zhang. Cosmology constraints from shear peak statistics in Dark Energy Survey Science Verification data. *ArXiv e-prints*, March 2016.

[159] J. Clampitt, C. Sánchez, J. Kwan, E. Krause, N. MacCrann, Y. Park, M. A. Troxel, B. Jain, E. Rozo, E. S. Rykoff, R. H. Wechsler, J. Blazek, C. Bonnett, M. Crocce, Y. Fang, E. Gaztanaga, D. Gruen, M. Jarvis, R. Miquel, J. Prat, A. J. Ross, E. Sheldon, J. Zuntz, T. M. C. Abbott, F. B. Abdalla, R. Armstrong, M. R. Becker, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, J. Estrada, A. E. Evrard, A. Fausti Neto, B. Flaugher, P. Fosalba, J. Frieman,

R. A. Gruendl, K. Honscheid, D. J. James, K. Kuehn, N. Kuropatkin, O. Lahav, M. Lima, M. March, J. L. Marshall, P. Martini, P. Melchior, J. J. Mohr, R. C. Nichol, B. Nord, A. A. Plazas, A. K. Romer, E. Sanchez, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, V. Vikram, and A. R. Walker. Galaxy-Galaxy Lensing in the DES Science Verification Data. *ArXiv e-prints*, March 2016.

[160] E. J. Baxter, J. Clampitt, T. Giannantonio, S. Dodelson, B. Jain, D. Huterer, L. E. Bleem, T. M. Crawford, G. Efstathiou, P. Fosalba, D. Kirk, J. Kwan, C. Sánchez, K. T. Story, M. A. Troxel, T. M. C. Abbott, F. B. Abdalla, R. Armstrong, A. Benoit-Lévy, B. A. Benson, G. M. Bernstein, R. A. Bernstein, E. Bertin, D. Brooks, J. E. Carlstrom, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, R. Chown, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. E. Evrard, A. Fausti Neto, B. Flaugher, J. Frieman, D. Gruen, R. A. Gruendl, G. Gutierrez, T. de Haan, G. P. Holder, K. Honscheid, Z. Hou, D. J. James, K. Kuehn, N. Kuropatkin, M. Lima, M. March, J. L. Marshall, P. Martini, P. Melchior, C. J. Miller, R. Miquel, J. J. Mohr, B. Nord, Y. Omori, A. A. Plazas, C. L. Reichardt, A. K. Romer, E. S. Rykoff, E. Sanchez, I. Sevilla-Noarbe, E. Sheldon, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, A. A. Stark, M. E. C. Swanson, G. Tarle, D. Thomas, A. R. Walker, and R. H. Wechsler. Joint Measurement of Lensing-Galaxy Correlations Using SPT and DES SV Data. *ArXiv e-prints*, February 2016.

[161] E. S. Rykoff, E. Rozo, D. Hollowood, A. Bermeo-Hernandez, T. Jeltema, J. Mayers, A. K. Romer, P. Rooney, A. Saro, C. Vergara Cervantes, H. Wilcox, T. M. C. Abbott, F. B. Abdalla, S. Allam, J. Annis, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. Brooks, D. L. Burke, D. Capozzi, A. Carnero Rosell, M. Carrasco Kind, F. J. Castander, M. Childress, C. A. Collins, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, T. M. Davis, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. E. Evrard, D. A. Finley, B. Flaugher, P. Fosalba, J. Frieman, K. Glazebrook, D. A. Goldstein, D. Gruen, R. A. Gruendl, G. Gutierrez, M. Hilton, K. Honscheid, B. Hoyle, D. J. James, S. T. Kay, K. Kuehn,

N. Kuropatkin, O. Lahav, G. F. Lewis, C. Lidman, M. Lima, M. A. G. Maia, R. G. Mann, J. L. Marshall, P. Martini, P. Melchior, C. J. Miller, R. Miquel, J. J. Mohr, R. C. Nichol, B. Nord, R. Ogando, A. A. Plazas, K. Reil, M. Sahlén, E. Sanchez, B. Santiago, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, J. P. Stott, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, D. Tucker, P. T. P. Viana, V. Vikram, A. R. Walker, and Y. Zhang. The redMaPPer Galaxy Cluster Catalog From DES Science Verification Data. *ArXiv e-prints*, January 2016.

[162] C. Chang, A. Pujol, E. Gaztañaga, A. Amara, A. Réfrégier, D. Bacon, M. R. Becker, C. Bonnett, J. Carretero, F. J. Castander, M. Crocce, P. Fosalba, T. Giannantonio, W. Hartley, M. Jarvis, T. Kacprzak, A. J. Ross, E. Sheldon, M. A. Troxel, V. Vikram, J. Zuntz, T. M. C. Abbott, F. B. Abdalla, S. Allam, J. Annis, A. Benoit-Lévy, E. Bertin, D. Brooks, E. Buckley-Geer, D. L. Burke, D. Capozzi, A. C. Rosell, M. C. Kind, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, T. F. Eifler, J. Estrada, A. E. Evrard, B. Flaugher, J. Frieman, D. A. Goldstein, D. Gruen, R. A. Gruendl, G. Gutierrez, K. Honscheid, B. Jain, D. J. James, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, M. Lima, J. L. Marshall, P. Martini, P. Melchior, C. J. Miller, R. Miquel, J. J. Mohr, R. C. Nichol, B. Nord, R. Ogando, A. A. Plazas, K. Reil, A. K. Romer, A. Roodman, E. S. Rykoff, E. Sanchez, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, and A. R. Walker. Galaxy bias from the DES Science Verification data: combining galaxy density maps and weak lensing maps. MNRAS, April 2016.

[163] M. Crocce and et al. Optimisation of the galaxy sample from Y1-DES data for BAO analysis. in prep.

[164] P. Fosalba, E. Gaztañaga, F. J. Castander, and M. Manera. The onion universe: all sky lightcone simulations in spherical shells. MNRAS, 391:435–446, November 2008.

[165] N. Benítez. Bayesian Photometric Redshift Estimation. ApJ, 536:571–583, June 2000.

[166] N. Benítez, H. Ford, R. Bouwens, F. Menanteau, J. Blakeslee, C. Gronwall, G. Illingworth, G. Meurer, T. J. Broadhurst, M. Clampin, M. Franx, G. F. Hartig, D. Magee, M. Sirianni, D. R. Ardila, F. Bartko, R. A. Brown, C. J. Burrows, E. S. Cheng, N. J. G. Cross, P. D. Feldman, D. A. Golimowski, L. Infante, R. A. Kimble, J. E. Krist, M. P. Lesser, Z. Levay, A. R. Martel, G. K. Miley, M. Postman, P. Rosati, W. B. Sparks, H. D. Tran, Z. I. Tsvetanov, R. L. White, and W. Zheng. Faint Galaxies in Deep Advanced Camera for Surveys Observations. ApJS, 150:1–18, January 2004.

[167] G. Favole, J. Comparat, F. Prada, G. Yepes, E. Jullo, A. Niemiec, J.-P. Kneib, S. A. Rodríguez-Torres, A. Klypin, R. A. Skibba, C. K. McBride, D. J. Eisenstein, D. J. Schlegel, S. E. Nuza, C.-H. Chuang, T. Delubac, C. Yèche, and D. P. Schneider. Clustering properties of g-selected galaxies at $z \sim 0.8$. *ArXiv e-prints*, July 2015.

[168] A. Cooray and R. Sheth. Halo models of large scale structure. Phys. Rep., 372:1–129, December 2002.

[169] A. Klypin, G. Yepes, S. Gottlöber, F. Prada, and S. Heß. MultiDark simulations: the story of dark matter halo concentrations and density profiles. MNRAS, 457:4340–4359, April 2016.

[170] G. L. Bryan and M. L. Norman. Statistical Properties of X-Ray Clusters: Analytic and Numerical Comparisons. ApJ, 495:80–99, March 1998.

[171] R. K. Sheth and A. Diaferio. Peculiar velocities of galaxies and clusters. MNRAS, 322:901–917, April 2001.

[172] P. S. Behroozi, C. Conroy, and R. H. Wechsler. A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for 0 ¡ z ¡ 4. ApJ, 717:379–403, July 2010.

[173] R. M. Reddick, R. H. Wechsler, J. L. Tinker, and P. S. Behroozi. The Connection between Galaxies and Dark Matter Structures in the Local Universe. ApJ, 771:30, July 2013.

[174] M. Crocce, A. Cabré, and E. Gaztañaga. Modelling the angular correlation function and its full covariance in photometric galaxy surveys. MNRAS, 414:329–349, June 2011.

[175] E. Sanchez and et al. Extracting BAO from a photometric galaxy survey: comparison of methods. in prep.

[176] J. Ross, A. and et al. Optimizing BAO Measurements for photoz surveys: Application to Dark Energy Survey Galaxy Clustering. in prep.